

AKIVA & ISAAK MOISEEVICH YAGLOM

**PROBABILITY
AND
INFORMATION**

Contents

PREFACE TO THE FIRST RUSSIAN EDITION	vii
PREFACE TO THE SECOND RUSSIAN EDITION	xi
PREFACE TO THE THIRD RUSSIAN EDITION	xv
PREFACE TO THE ENGLISH EDITION	xix
 CHAPTER 1 Probability	
1.1 Definition of Probability. Random Events and Random Variables	1
1.2 Properties of Probability. Addition and Multiplication of Events. Incompatible and Independent Events	7
1.3 Conditional Probability	20
1.4 The Variance of a Random Variable. Chebyshev's Inequality and the Law of Large Numbers	26
1.5 Algebra of Events and General Definition of Probability	36
 CHAPTER 2 Entropy and Information	
2.1 Entropy as a Measure of the Amount of Uncertainty	44
2.2 The Entropy of Compound Events. Conditional Entropy	59
2.3 The Concept of Information	73
2.4 Entropy (revisited). The Determination of Entropy from its Properties	93
 CHAPTER 3 The Solution of Certain Logical Problems by Calculating Information	
3.1 Simple Examples	101
3.2 The Counterfeit Coin Problem	108
3.3 Discussion	121

CHAPTER	4	Application of Information Theory to the Problem of the Information Transmission Through Communication Channels	
	4.1	Basic Concepts. Efficiency of a Code	137
	4.2	Shannon-Fano and Huffman Codes. Fundamental Coding Theorem	147
	4.3	Entropy and Information of Various Messages Encountered in Practice	177
	4.4	Transmission of Information over Noisy Channels	258
	4.5	Error-Detecting and Error-Correcting Codes	304
APPENDIX	I	Properties of Convex Functions	347
APPENDIX	II	Some Algebraic Concepts	364
APPENDIX	III	Table of Values of $-p \log p$	392
APPENDIX	IV	Short Table of the Function	
		$h(p) = -p \log p - (1 - p) \log (1 - p)$	395
		References	397
		Name Index	409
		Subject Index	413

Preface to the First Russian Edition

For a long time it so happened that almost no information on the scientific research carried out in the field of mathematical theory percolated beyond the realm of a restricted circle of professional mathematicians. This circumstance sometimes even led the non-specialists to an entire incorrect notion of absolute completeness of mathematics, envisaging new research in this field to be almost impossible or, in any case, extremely tedious. The reason for such situation is explained by the fact that an overwhelming majority of recent works published in mathematical journals are related to sufficiently developed branches of this science which are incomprehensible to a person having no special training. As regards more elementary areas of mathematics, like elementary geometry, it is difficult to suppose the existence of any facts or theorems of really crucial theoretical value that has gone unnoticed by several generations of workers in this area.[†] Also the new significant directions that have emerged in pure and applied mathematics during the recent decades, as a rule, are confined to sufficiently complex concepts and ideas offering little scope for their popularization. Viewed in this context, the credit to C. E. Shannon, the well-known American applied mathematician, becomes all the more due, for his ability to inaugurate in 1947-1948 a new important domain of mathematics, which stemmed from quite elementary considerations.

The basic problems confronted by Shannon in the initiation of the discipline which was later designated as information theory, were connected with engineering questions related to electrical and radio communications.^{††} Generally speaking, newly emerging applications of mathematics in engineering and natural

[†]However, as a matter of fact, even in these elementary areas of mathematics some serious questions still remain open. Therefore, it is not surprising that sometimes stimulating and fundamental works appear related e.g., to elementary geometry. See, for instance, W.G. Boltvanskii's *Equivalent and Equidecomposable Figures*, Fizmatgiz, Moscow, 1956 (English translation published by D.C. Heath, Boston, 1963), based mainly on quite recent results from elementary geometry,

^{††}Owing to the general character of Shannon's work, it exerted a great stimulating effect on the entire research related to the transmission and preservation of any information met with in nature and technology. The channels through which this information is transmitted may be not only the telegraphic and telephonic wires or media transmitting radio-signals but also the nerves through which signals from organs of sense are transmitted to muscles via brain, or those yet almost completely unexplored paths by which the indications of future structural plan of living organism from an embryonic cell are transmitted.

sciences are usually closely related to the use of complex mathematical notions and methods. Hence quite often they are also not susceptible to elucidation without a deep insight into the intricate problems of modern science and technology. This circumstance has severely restricted the opportunities of popularization of recent practical achievements of mathematics. Hence, the idea of a non-specialist about the importance of applied mathematics often remains confined to his intelligence drawn from school courses regarding the fact that geometry was used in ancient Egypt for reestablishment of land boundaries after floods in the Nile and a few similar facts. And, in this respect, the exposition of a string of ideas related to the information theory represents an extremely alluring theme for popularization, since the simplest practical applications of these ideas to modern engineering problems can be explained fully even to the readers who have a minimum mathematical and engineering background.

The present book, designed for a wide circle of readers (familiarity with mathematics up to high school level suffices for comprehension of all of its contents), makes, of course, no claim to serve even as an elementary introduction to the scientific information theory. We can give here only a preliminary idea of important practical applications of this theory. Similarly, it shall not be possible to deal here with the deeper purely mathematical problems connected with the information theory. The main aim of the authors is much simpler: it consists of acquainting the reader with certain, though not complex but highly important, new mathematical ideas, and leading him through these ideas to an understanding of one of the possible means of employing mathematical methods of modern engineering.

The first chapter of the book is devoted to the exposition of the classical (introduced as early as seventeenth century) concept of *probability*, acquaintance with which is necessary for a comprehension of all the content matter that follows. In the second chapter, the recent concepts of *entropy* and *information* due to Shannon are considered, whose general scientific value has been evaluated by the mathematicians only during the last few years. The third and fourth chapters present examples and applications. In contrast to the preceding two chapters, rigorous proofs of statements made here are often just outlined or completely omitted, and in certain cases such statements are even formulated only in the form of highly plausible propositions. Furthermore, in the third chapter we have demonstrated the usefulness of the concepts of entropy and information by recreative problems on guessing numbers, counterfeit coins and so on (these problems are in a sense similar to problems on playing cards and dices that led to the emergence of probability theory in the seventeenth century). The engineering applications to communication theory that are richer in content are discussed in the fourth chapter. We expect the reader's acquaintance with the recreative third chapter to enable him to develop a better grasp of the meaning of basic concepts introduced in Chapter 2, and by the same token to prepare himself for a study of Chapter 4, which is the most complex part of the book

and also uses some results of the third chapter.

Though the book is designed for all lovers of mathematics, it is meant primarily for high school and undergraduate college students and teachers. Together with them, it must also be of interest to many readers who have specialized in communication engineering but do not possess a sound mathematical background. The book is based on a lecture delivered by one of the authors to a group of high school participants of a special mathematics study group at Moscow State University. The contents of this lecture have, however, been expanded considerably here.

The authors express their sincere gratitude to A. N. Kolmogorov, whose valuable suggestions contributed to an appreciable improvement of the book. They are also thankful to M. M. Goryachaya, the editor of the book, whose remarks helped in correcting certain deficiencies of the primary exposition.

Moscow, May, 1956

A. M. YAGLOM

I. M. YAGLOM

Preface to the Second Russian Edition

The second edition of the book *Probability and Information* does not differ in structure substantially from the first edition. A comparison of the table of contents of both the editions of the book will make it clear that the structural variations between the two are quite insignificant. The character of the book has also not been changed, assuming of the reader a quite modest mathematical knowledge (which deficiency must, however, be counterbalanced with certain persistence). Nevertheless, there are specific distinctions between the two editions which are so significant that we may now speak of it as being a new book rather than a revised edition.

Such crucial changes have partially stemmed from the fact that this book deals with a very young and rapidly developing branch of science, for which an interval of two years between the first and the second editions constitutes a noteworthy gap. The authors tried to keep themselves abreast with the developments that took place during these two years. This was accomplished by them to a great extent by looking over numerous new books and papers, since the literature on information theory has proliferated during this period with stupendous intensity. However, it is the one omission due to the authors which singularly necessitated the revision of the first edition.

The present book has grown from a lecture delivered to a group of Moscow high school students interested in mathematics. The authors firmly bear in mind this genesis of the book, to which the readers obviously pay little attention. Accordingly, in the Preface to the first edition of the book it was stated that it is designed for all lovers of mathematics and primarily for high school teachers and students. In this connection, we, however, overlooked one more category of numerous readers, consisting of people who are seriously interested in the information theory (and not in mathematics in general), but do not desire to embark upon its study through specialized literature, whose thorough grasp involves both time and efforts. The book drew the greatest appreciation from the professional mathematicians and communication engineers and our remonstrances that it was not intended for readers from either of these categories failed to produce any effect. The authors were taken by surprise by the swiftness with which the first edition of the book disappeared from the market and was translated in several foreign languages (e.g., Hungarian, German, French and Japanese). Such overwhelming response forced us to concede that the book does meet some vital needs and prompted us to focus our attention on how this requirement could be served more adequately.

We are now also inclined to consider that our book is unsuitable for the readers who are interested in the sophisticated topics of the mathematical information theory or of communication engineering. For the former class of readers, it is natural to recommend Feinstein [9]†, comparatively a concise but terse book. For readers of the second category, Woodward [24] is obviously a quite suitable and fascinating work. Also, the physicists or biologists, who are interested in Shannon's ideas, would not naturally turn to our book but to Brillouin [5] (for physicists) and Ashby [3] (for biologists). However, it could possibly be profitable even for many readers from all such categories to acquaint themselves with the present elementary book as a starting point. It is only for the philologists, who currently represent a sufficiently significant group of 'users' of information theory that we had nothing to suggest; this led us to devote greater attention to the problems encountered by them in the second edition of our book. And, if during the preparation of the new edition we have rejected as before any material whose inclusion could raise the mathematical level beyond what is required for the reading of the first edition, then we have also kept in view this time not only the school students, but also the biologists or philologists who are not familiar with calculus.

This requirement to cater for a wider circle of readers of the book necessitated a series of essential changes in the text. Thus, for example, in the new edition the capital Russian letters Ξ (entropy) and \mathbb{V} (information) are removed. In fact, these unusual notations could have facilitated the reading of the book by completely inexperienced readers, but at the same time they caused inconvenience to all those who had (or desired in future to have) to do also with other literature on information theory, using different notations. It was also natural that in Chapter 2 we paid adequate attention to the statistical interpretation of the concept of entropy, making it quite fruitful for all practical applications of information theory. We have considerably expanded the last chapter that has the greatest applied value; the volume of the book has also been enlarged with the addition of the supplementary material printed in small type (that may even be skipped in the first reading). In particular, taking account of the interest of mathematicians, we have derived in these supplements rigorous proofs of certain premises that have been merely propounded in the basic text. The character of the problems in the book has also been changed; in the present edition exercises on the urn scheme and mathematical recreations occur uncommonly but then the practical problems of the applied information theory are more frequent. However, we have preserved the entire chapter devoted especially to the elementary problems on quick wits, since it is essentially through these problems in a new (and sufficiently attractive) form that we have tackled pretty serious questions of the most economic message transmission. This rela-

†The digits in the square bracket indicate the number at which the reference is listed at the end of the book.

tionship which we found to have been missed by some readers of the first edition, has now been more prominently highlighted.

The present edition of the book is supplemented with a bibliography which the first edition had lacked. Being convinced (in particular by the experience gained during our work on this book) of the computational convenience that is conferred by the table of values of the function $-p \log p$ (where $0 \leq p \leq 1$), we have included such table as Appendix III in the book. The binary system of logarithms has been retained in this table; in the text, however, we have employed decimal logarithms, which have a wider acceptance from the majority of readers (especially because we desired to demolish the notion held by several engineers that the use of binary logarithms forms precisely the basis of information theory).

In conclusion, the most significant change is the addition of a special Section 4.3, which gives a resume of the data on information contained in various specific types of messages (written and spoken language, music, television and phototelegraphic images). At the end of this section we have also briefly cited some data on the capacity of different communication channels. This is the largest section in the book; it is not used directly in the following text and can be skipped completely by a reader who is interested in only the mathematical side of information theory. To us, however, it appears that the number of those readers will be considerably large for whom this section proves to be of highest appeal. Section 4.3 is somewhat of a distinctive character from the rest of the book—factually, it presents a review of a large number of comparatively more specialized papers that have appeared recently in different scientific and engineering journals. For the convenience of readers who are interested in some specific field of the applications of information theory, we have indicated in all cases the exact source that contains a more elaborate exposition of the results mentioned by us (the major portion of the bibliography appended to the book is related to this section). It has also been our endeavour to make our review as complete as possible (to the extent to which it could be possible without violating the elementary character of the book). However, it has been necessary to bear in mind that owing to the intensity with which the study of statistical properties of messages and communication channels is being pursued all over the world during present times, it is apprehended that the review Section 4.3 may become deficient by the time the book appears and a few years later the data presented in it may become substantially outdated. However, we feel that even then Section 4.3 shall not lose its utility. In fact, the basic objective of this section is to give an idea of the order of magnitudes of the amount of information met with in science and technology, and to illustrate the general directions in which engineering, philological and biological studies have been inspired by information theory, but not to provide at all a base for the scientific research work of specialists.

Finally, we wish to thank sincerely all our readers who communicated us their comments, which assisted us in the preparation of the new improved

edition. In particular, we wish to thank S. G. Gindikin, A. N. Kolmogorov, V. I. Levenstein, P. S. Novikov, I. A. Ovseevich, S. M. Rytov, V. A. Uspenski, G. A. Shestopal, M. I. Eidelnant and especially R. L. Dobrushin and A. A. Kharkevich. We are also grateful to V. A. Garmash, L. R. Zinder, D. S. Lebedev and T. N. Moloshnaya for fruitful discussions we had with them on the problems connected with the contents of Section 4.3 of the present book.

Moscow, March 1959

A. M. YAGLOM

I. M. YAGLOM

Preface to the Third Russian Edition

The first edition of the book was published in 1957 and the second one in 1960. However, there is a passage of thirteen years between the second and third editions. We ourselves must apologize for such a considerable gap between the last two editions. Though the second edition of the book was long back reduced to the status of a mere bibliographic rarity leading to a spate of enquiries from the readers and repeated overtures from the publishing house for its revision, we could not somehow make up our mind. It was clear to us that it was impossible to keep the book in the form it had in the second edition, because it was necessary to incorporate in it the substantial changes that had taken place during these years in information theory. Such thorough revision of the book (accompanied with the alteration of even its title as suggested by many) obviously presented a highly laborious and involved task, which was perhaps beyond our stamina.

We eventually took recourse to the way of compromise, which is almost always chosen by the people placed in an inconvenient situation. The present third edition of the book retains the earlier title and much of its original look. Thus, for example, we do not assume of the reader, as before, the background beyond the level of high school mathematics. The book accordingly still remains simpler than all the other existing text books and monographs giving an exposition of information theory. At the same time, we could not also ignore the circumstance that, to our surprise, the second edition of *Probability and Information* was used both within and outside our country in a series of cases as the basic textbook for delivering lecture courses in colleges and universities. Hence during the revision and enlargement of the text we had the added impetus to make the book more suitable for such use, earlier not foreseen by us. In particular, we have refrained from using in the text the common decimal logarithms and the uncommon decimal units for the measurement of the amount of information (dits) which thereby eliminated the last shred of direct evidence of this book having grown out of a lecture delivered to school students many years ago.†

The last Chapter 4, which is also the most important chapter in the book, has

†In the literature addressed to school students, the use of binary logarithms creates some impression of artificiality. However, in a book on information theory designed for more mature readers, such impact is liable, contrarily, to promote the employment of decimal logarithms in place of the universally used binary ones,

undergone the maximum revision, since Chapters 1—3 actually represent only an introduction to the basic content material of the book that has been brought into focus in Chapter 4. Keeping in view the readers who desired to be acquainted through this book with the mathematical fundamentals of information theory, we have included in Section 4.2 an exposition of optimal Huffman codes (theoretically more important than the Shannon-Fano code considered in the previous editions also) and substantially sharpened the proof of fundamental noiseless coding theorem, making it more compact and mathematically precise. Section 4.4 is still more extensively modified where we have deduced, in particular, two new proofs of the fundamental noisy coding theorem together with a simple proof of the converse to coding theorem. The same purpose is also served by the inclusion in first chapter of the law of large numbers, which permits us to make later some more rigorous deductions, and also by an appreciable increase in the number of references from serious scientific literature, to the study of which this book provides a natural bridge.

However, the most crucial circumstance we had to take into consideration in the preparation of the revised edition of our book is that during the last two decades even the frontiers of information theory underwent a substantial change. In the present times, the most important part of information theory is indisputably the *coding theory*, whose rapid development was impossible to be forecast at the time the earlier editions were written. Hence, today even a popular work on information theory will be irrelevant if it completely ignores that branch of this subject which attracts greatest interest of both the theoreticians and practical engineers and engages lion's share of efforts of the specialists in information theory throughout the world. On the other hand, the general character of coding theory and mathematical tools and methods applicable to this important and elegant field of applied mathematics differ quite substantially from the basic contents of our book. The reorientation of the book to the direction of coding theory would have involved rewriting the book afresh. Hence, here also we have kept to the middle of road: we have added to Chap. 4 a completely new concluding section to provide just an introduction to the tasks and techniques of coding theory; as a matter of fact, even in its present form this section is appreciably out of tune with the rest of the contents of the book. This gap motivated us to add to the book a new Appendix II devoted to certain purely algebraic concepts and propositions; however, as a compensatory feature we have omitted Appendix II of the second edition as it had become superfluous after the revision of the main text. Strictly speaking, the new Appendix II is not prerequisite for following the content matter of Section 4.5 devoted to the coding theory; however, an overview of this appendix before taking up the indicated section will obviously enable the reader to have a greater insight into the potentialities of further development and extension of the results of this section.

A singular place is occupied in this book by Section 4.3 of which we have said in sufficient length in the Preface to the second edition. This section contains

a review of the data on various specific types of messages which as far as known to us is a unique resume of this kind in the literature; the latter circumstance also motivated us to enlarge this section further by including in it a review of the majority of more recent works. It is obvious that in spite of the extensive expansion of the reference list, we cannot make a pretence to have covered all of the printed literature on the topics considered. It is quite possible that works scattered over a vast number of journals in highly diverse stray fields might have escaped our notice. We must also caution the reader that we have not concerned ourselves with the verification of numerical data available in various investigations and an analysis of the extent of their statistical reliability. It seems that much work still remains to be done in the latter direction. However, despite the fact that not all of the data adduced in Section 4.3 is completely reliable, its inclusion in the book is justified, for it enables the reader to get here a sufficiently complete idea of the results achieved so far in a number of specific fields of information theory and of the general directions of major researches in these fields.

Of course, many aspects related to information theory have not been touched upon in our book. Apart from the natural infeasibility 'to envelop the boundless', the limitation is partly set upon by our proneness to retain in the present edition the look this book earlier had. Thus, for example, we have as before almost completely ignored in it the problems connected with the estimation of entropy and information of experiments with an *infinite* set of possible outcomes (as regards the general concepts and definitions involved here see, for example, [12]). We do not also concern ourselves with the so-called 'algorithmic' approach to the concept of the amount of information (for salient works in this direction, see, for example, [15] and [27]); moreover, a combinatorial treatment of this concept is only briefly sketched in Section 4.3. Finally, all attempts at broad interpretations of the concept of information beyond the framework of Shannon's theory (of the type of 'semantic information' or 'thesarus'; see, for example, [4], [13] and [20]) fall beyond the scope of this book (these attempts are of quite preliminary nature till now).

As is well known, the main value of Preface is that it enables the authors to thank all those who assisted them in their work. A. N. Kolmogorov has been kind enough to place at our disposal his remarkable (unpublished) manuscript designed to refine substantially Shannon's guessing method for estimation of the entropy of written language which has been discussed at length in this book. Some additional material related to the entropy of a language has also been contributed by A. V. Prokhorov. We must also mention that V. V. Ivanov, I. A. Ovseevich, N. V. Petrova, B. S. Tsybakov and W. Endres brought to our notice some literary sources, which we used to enlarge Section 4.3. The contents at a number of places in the book bear the stamp of numerous discussions we had with R. L. Dobrushin on the topics from information theory. S. Z. Stambler, the editor of the third edition, carefully read the entire text and contributed

to its further improvement. He also supplied us with a long list of additional references that we have used during the preparation of our book. We express our sincere gratitude to all the persons mentioned here.

Moscow, March 1972

A. M. YAGLOM

I. M. YAGLOM

Preface to the English Edition

The panoramic history of this book is described in the prefaces of its Russian editions. It has taken to a chequered and not customary course of development; to start with it was a small elementary book for teenagers based on a lecture delivered by one of us 23 years ago to a group of Moscow high school students. The primary aim of the book was to expose the relationship of certain mathematical recreation exercises with rather serious and very interesting mathematical methods developed recently in engineering sciences in order to stimulate young readers' interest in modern mathematics. Later, however, the book began to live independently of our wishes. We received a lot of letters and comments from our readers and almost all of them turned out to be grown-ups having no leisure for recreations but seriously interested in the information theory. Therefore, we changed considerably the scope of our book in the second and third editions in an effort to meet the demands of the new (and, as we discovered, the predominant) category of our readers. As a result, the book developed into a thick volume intended for a wide community of people interested in various applications of the modern information theory, but having no special mathematical background (in fact, even the requirement of the knowledge of elementary differential calculus is dispensed with in our book).

The book scored a remarkable success in other countries also, and this was obviously caused by a widespread interest in the ideas of information theory all over the world. The book was translated into at least 10 foreign languages and some of the translations underwent several editions which differed from each other (and also from all the corresponding Russian editions since, wherever possible, we tried to send to the publishers some supplementary material). However, for a long time the opportunity of the publication of English translation kept on eluding us. We received twice letters from publishers of repute (one in the U.S.A. and the other in U.K.), seeking our permission to publish the English edition of the book. In both the cases, we gave the permission and even sent some corrections and supplements. It seems to us that on both the occasions the translation work was started but then some technical difficulties thwarted the completion of the work. Therefore, we are happy that Hindustan Publishing Corporation have finally published the English translation of our book and thus made it accessible to a wide circle of new readers.

The English edition differs from all the previous ones. Besides the minor corrections and improvements, we have completely revised (and considerably

extended) Section 4.3, for it is clear that the discussion of the amount of information contained in the spoken and written text messages must now be based on the data related to the English (and not Russian) language. We have also enlarged the concluding Section 4.5 by supplementing it with a description of the method of constructing the practically important Bose-Chaudhuri-Hocquenghem error-correcting codes. This necessitated the inclusion of some additional material in Appendix II at the end of the book, since the role of this appendix was widened further in comparison to the Russian edition. In order to update the book, Section 4.3 has been further reinforced by inclusion of the description of some latest works though, of course, it is not possible to claim that we have covered all recent papers, which are too numerous to cater for. We have also added a new Appendix IV which contains a short table of the function $h(p) = -p \log p - (1 - p) \log (1 - p)$, keeping in view the usefulness of such table for educational purposes, which is one of the avowed objectives of the book.

We are glad to express here our appreciation to Hindustan Publishing Corporation for production of the book and to Drs. B. Mandelbrot, T. M. Cover and T. Nemetz who have sent us some new material used in the preparation of the present edition.

Moscow and Yaroslavl
June, 1983

A. M. YAGLOM
I. M. YAGLOM

1

Probability

1.1. Definition of probability. Random events and random variables

In practice, we quite frequently encounter experiments (variously, trials, observations, processes) which yield different results depending on whether the situations are unknown or unaccounted for. Thus, for example, when throwing a die (a homogeneous cube with its faces numbered from 1 to 6), we cannot know in advance what face will turn up, since this depends on many unknown factors (details of hand movement resulting in throwing, die position at the instant of roll, peculiarities of the underlying surface, and so on). It is equally impossible to forecast beforehand the number of secondary school graduates that will enter a given college during a specific year, the number of defective items produced by a factory on a given day, or the number of rainy days that will occur next year. Similarly, there is no way of predicting the number of errors that will be committed by a school student in the homework, or the ticket number that will draw the first prize in a prospective lottery draw (the number of winning tickets are determined by drawing from a well-shuffled lot of numbered tickets in a container), and so on. The number of similar examples can obviously be augmented considerably.

The application of mathematics to a study of such phenomena is based on the following fact. In many cases when the same experiment is repeated many times under identical conditions *the frequency of occurrence of the result* under consideration (i.e., the ratio of the number of occurrences of this result to the total number of trials) *always remains approximately the same*, close to some constant number p . For example, it is thus known that the frequency with which a gun will hit a target under a given set of shooting conditions, as a rule, always remains almost the same and seldom deviates significantly from a certain average number (with the passage of time, this average number may apparently vary—in such cases we say that the marksman is improving upon, or conversely worsening his performance). Also the frequency with which a six shows up on the die or the percentage of defective items under a given set of conditions usually deviates little when the related ‘trials’ (throw of die or manufacture of a given item) are repeated on a mass scale. Proceeding from this, we conclude that in each case there exists a definite constant number which objectively characterizes the very process of shooting, die rolling, production of items, and so on. About this

constant the average frequency for the corresponding outcome (hits of a target, appearance of a six, emergence of defective items) fluctuates all the time (but does not deviate from it significantly) in the given series of 'trials.' The corresponding constant number is called *probability* of the event under investigation. Probability is defined similarly in a series of other problems related to such widely divergent fields as mathematics, mechanics, physics, engineering, economy and biology. The discipline that studies the properties of probability and various applications of this concept is called the *probability theory*.

According to the discussion above, the probability of some event can be evaluated approximately from the outcomes of a long series of trials. However, obviously the very existence of a probability does not depend ultimately upon whether an experiment is performed or not. This raises a most natural question concerning the methods by which one can compute the probabilities of various events without first carrying out the corresponding experiments; by applying such methods we can make, beforehand, a forecast about the outcome of a succeeding trial, thus opening up great opportunities for the practical scientific applications of the concept of probability. We shall not undertake here a detailed discussion of this question, but shall confine ourselves only to a very simple example from which, however, there can be derived a comparatively wide range of problems concerning the evaluation of probability.†

Suppose that we have a box (or as it is often said an urn) containing 10 well-mixed balls distinguishable from each other only by colour. Of these 5 are white, 3 black, and 2 red. We draw a ball from the urn without looking at it; the question is: What is the probability that this drawing will produce a ball of a specific colour? It is perfectly clear that here the chances are that out of 10 drawings 5 will produce a white ball, 3 a black ball, and 2 red one; in other words, the probability of drawing a white, black, or red ball is, respectively, $\frac{5}{10} = \frac{1}{2}$, $\frac{3}{10}$, and $\frac{2}{10} = \frac{1}{5}$. Also, indeed, if we repeat this particular experiment many times (every time returning the ball drawn to the urn and mixing all the balls well, we become convinced that, of all the drawings, roughly 50% result in a white ball, 30% in a black ball, and 20% in a red one. Naturally, the problem of determining the probability of any other configuration of balls of diverse colours, well-mixed and contained in an urn, is also solved in the same straightforward manner.

Let us consider a few more problems of the determination of probability, which reduce to the 'urn model.'

†The book by B. V. Gnedenko and A. Ya. Khinchin [31] is recommended to the reader desirous of a more thorough study of the probability theory and its applications. A much bigger but quite readable book by P. Mosteller, R. E. K. Rourke and G. B. Thomas [38] is also highly suitable for primary acquaintance with the probability theory. See also, slightly more difficult articles by A. N. Kolmogorov [35] and M. Kac [33] and other related references at the end of this book.

Problem 1. *In flipping a coin at random, what is the probability that a 'head' will show up?*

This problem is obviously equivalent to the scheme of placing two balls in an urn, of which one is marked 'head' and the other 'tail' (of course, instead of inscribed balls, one can consider balls of two different colours, for example, a white and a black). What is the probability that a random drawing of a ball from the urn will produce a ball inscribed 'head'? It is clear that the desired probability here is $\frac{1}{2}$.

Problem 2. *In rolling a die at random, what is the probability of getting an integer divisible by 3?*

Instead of rolling a die, we may speak of drawing a ball from an urn containing six balls numbered 1, 2, 3, 4, 5, and 6. Now if the third and sixth balls are coloured black and the rest are left white, we arrive at the problem of determining the probability of drawing a black ball (the numbers 3 and 6 are divisible by 3, but the others are not). It is evident that the desired probability here is $\frac{2}{6} = \frac{1}{3}$.

Problem 3. *The gathering at a students' evening is known to consist of twenty students from the first college, twenty-five from the second, and thirty from the third college. What is the probability that the student with whom you randomly talked studies at the second college?*

This problem obviously corresponds to the scheme of an urn containing 75 balls, of which 20 are white, 25 are black, and 30 are red. What is the probability that when a ball is drawn randomly from the urn, it will be a *black* one? Clearly, this probability is $\frac{25}{75} = \frac{1}{3}$.

We now proceed to grasp the general principles of solving all these problems. In the urn scheme that we discussed as a preface to these problems, the condition that the balls in the urn be well-mixed and drawn blindly implies that we may, with equal justification, expect the appearance of any of the balls contained in the urn or, in other words, that the drawing of every ball is equally probable. But, since, there are in all 10 balls, it is natural to infer that the probability for a particular ball to be drawn is $\frac{1}{10}$. Further, since there are five white balls, the probability of drawing a white ball is $\frac{5}{10} = \frac{1}{2}$.

Exactly the same reasoning leads to the answers to Problems 1-3 above. Thus, for instance, in the case of a die roll, we assumed the appearance of any of the six faces to be equally probable; it is just this reason that enabled us to replace this problem by that of making a drawing from an urn containing six balls. However, of the six faces, precisely two are such that their appearance satisfies the hypothesis of the problem; the probability of the appearance of either of these two faces is $\frac{2}{6} = \frac{1}{3}$.

If we postulate that the experiments under consideration (drawing of a ball from an urn, tossing of a coin, rolling of a die, conversation with one of the

participants at a students' evening, etc.) have n equally probable outcomes, then it is necessary to regard each of these outcomes as having probability $1/n$. We now consider some event (the drawing of a white ball from an urn, occurrence of a 'head' when a coin is tossed, appearance of an even number when a die is rolled, a conversation with a student studying in the second college, and so on) to be determined by the outcomes of an experiment. If this event is realised in m out of all n equally probable outcomes of an experiment but not in the remaining $n - m$ outcomes, then the probability of its occurrence is taken as m/n . In other words, *the probability of a certain event is equal to the ratio of the number of equally probable outcomes favourable to the given event to the total number of equally probable outcomes*. The italicised matter may be taken as the definition of the concept of probability: further, it must be stipulated in the description of the experiment to be performed that the distinct outcomes are equally probable. This objective is precisely served by indicating that the die has the exact shape of a cube and is made of homogeneous material, or that the balls are well-mixed and are indistinguishable from each other (except with regard to colour). Although, such a definition does not cover some important cases of the evaluation of probability (see, for example, papers [33], [35] and books [29], [30] and [39] as well as Section 5 of this chapter printed in small type), it is adequate for a majority of the cases considered in this book.

Let us now agree on the terminology which we will need later on. An event which may or may not occur as the result of an experiment is called a *random event*; in the same sense we speak of the outcome of a given experiment. We shall use capital letters to denote random events and denote by p the probability of a random event (or of a specific outcome of an experiment); the probability of an event A is often written as $p(A)$. An important role is played by an experiment that can have *several* different outcomes; in such a case, we denote all these outcomes by a single letter with different subscripts (and the experiment itself mostly by a Greek letter).

To each such experiment there corresponds a specific probability table:

Outcomes of experiment	A_1	A_2	\dots	A_k
Probability	$p(A_1)$	$p(A_2)$	\dots	$p(A_k)$

Thus, for example, the urn experiment discussed on p. 2 corresponds to the table

A_1	A_2	A_3
$\frac{1}{2}$	$\frac{3}{10}$	$\frac{1}{5}$

(here, A_1 is the drawing of a white ball, A_2 of a black ball, and A_3 of a red ball). The experiment considered in Problem 1 is characterized by the simple

table :

B_1	B_2
$\frac{1}{2}$	$\frac{1}{2}$

(here, B_1 and B_2 represent, respectively, the 'head' and 'tail' that can appear). The rolling of a die gives the following probability table:

<i>Number that appears on the face</i>	1	2	3	4	5	6
<i>Probability</i>	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

It is worthwhile to note one salient difference between the last table and its two predecessors. The outcomes of the last experiment can be expressed by means of specific *numbers* (1, 2, 3, 4, 5 and 6), an opportunity not open to us in the

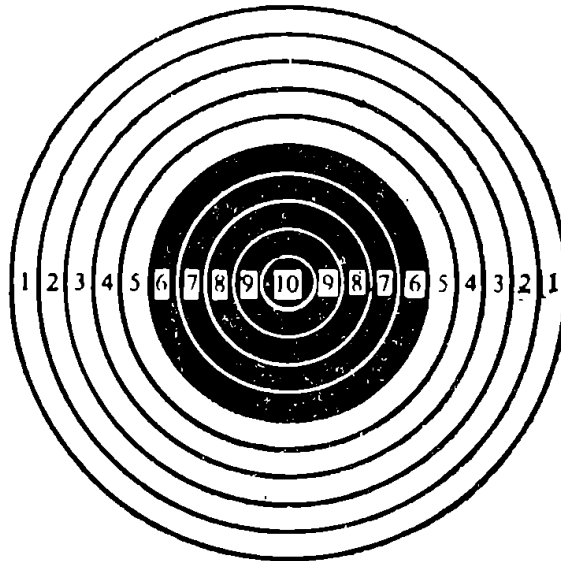


Fig. 1.

preceding examples. In this case, we can say that the number that appears on a face when a die is rolled, is a *random variable* which is capable of taking any one of all the six possible values, depending on chance (i.e., depending upon situations that are not subject to predictability). Other examples of random variables are the number of defective items per lot of 100, the number of births in some town per annum, the number of points scored by some marksman under prescribed shooting conditions in one round of firing (a target board showing

the number of points that are counted when each of its parts is hit is illustrated in Fig. 1), etc.†

The very term 'random variable' demands that its value may vary but nevertheless it may be somehow evaluated. The way this is to be achieved is not difficult to conceive. As an example, let us consider the first random variables enumerated above (the number of defective items in a lot of 100); suppose that, under defined production conditions, this number does not exceed 6, and the corresponding probabilities take the form of the following table:

<i>Number of defective items</i>	0	1	2	3	4	5	6
<i>Probability</i>	0.1	0.15	0.2	0.25	0.15	0.1	0.05

In such a case, in a large number N times a hundred items, roughly $0.1N$ do not contain a single defective item, $0.15N$ contain one, $0.2N$ two, $0.25N$ three, $0.15N$ four, $0.1N$ five and $0.05N$ six defective items. Consequently, for large N , the mean number a of defective items can be regarded as given by

$$a = 0.1N \cdot 0 + 0.15N \cdot 1 + 0.2N \cdot 2 + 0.25N \cdot 3 + 0.15N \cdot 4 + 0.1N \cdot 5 + 0.05N \cdot 6.$$

Hence, the mean value of the number of defective items per hundred (the mean percentage of defects) is given here by

$$a/N = 0.1 \cdot 0 + 0.15 \cdot 1 + 0.2 \cdot 2 + 0.25 \cdot 3 + 0.15 \cdot 4 + 0.1 \cdot 5 + 0.05 \cdot 6 = 2.7.$$

In general, if the probability table for the random variable α has the form

<i>Values of random variable</i>	a_1	a_2	a_3	\dots	a_k
<i>Probability</i>	p_1	p_2	p_3	\dots	p_k

then the *mean value* of this random variable is defined by

$$\text{m.v. } \alpha = p_1 a_1 + p_2 a_2 + p_3 a_3 + \dots + p_k a_k.$$

From this formula it follows, in particular, that the *mean value of a random variable* is just the *mean*, i.e., it *never exceeds its maximum possible value nor is less than its minimum value*. In fact, suppose that a_1 is the maximum value of the random variable α (i.e., $a_1 \geq a_2, a_1 \geq a_3, \dots, a_1 \geq a_k$) and a_k is its least

†The concept of random variables is incidental to the main theme of this book but occupies a central position in the theory of probability. In this connection, see, for example, the second part of B. V. Gnedenko and A. Ya. Khinchin's book [31].

value (i.e., $a_k \leq a_1, a_k \leq a_2, \dots, a_k \leq a_{k-1}$), then

$$\begin{aligned} \text{m.v. } \alpha &= p_1 a_1 + p_2 a_2 + \dots + p_k a_k \leq p_1 a_1 + p_2 a_1 + \dots + p_k a_1 \\ &= (p_1 + p_2 + \dots + p_k) a_1 = a_1, \end{aligned}$$

and

$$\begin{aligned} \text{m.v. } \alpha &= p_1 a_1 + p_2 a_2 + \dots + p_k a_k \geq p_1 a_k + p_2 a_k + \dots + p_k a_k \\ &= (p_1 + p_2 + \dots + p_k) a_k = a_k \end{aligned}$$

(for $p_1 + p_2 + \dots + p_k = 1$).

Problem 4. Suppose that the probability tables showing the frequency for marksmen *A* and *B* hitting the target have the form:

(i) For marksman *A*

Number of points	0	1	2	3	4	5	6	7	8	9	10
Probability	0.02	0.03	0.05	0.1	0.15	0.2	0.2	0.1	0.07	0.05	0.03

(ii) For marksman *B*

Number of points	0	1	2	3	4	5	6	7	8	9	10
Probability	0.01	0.01	0.04	0.1	0.25	0.3	0.18	0.05	0.03	0.02	0.01

Which of *A* and *B* should be regarded as the better marksman?

Here, the mean number of points scored by *A* in one round is

$$\begin{aligned} &0.02 \cdot 0 + 0.03 \cdot 1 + 0.05 \cdot 2 + 0.1 \cdot 3 + 0.15 \cdot 4 + 0.2 \cdot 5 + 0.2 \cdot 6 \\ &+ 0.1 \cdot 7 + 0.07 \cdot 8 + 0.05 \cdot 9 + 0.03 \cdot 10 = 5.24, \end{aligned}$$

and that for *B* is

$$\begin{aligned} &0.01 \cdot 0 + 0.01 \cdot 1 + 0.04 \cdot 2 + 0.1 \cdot 3 + 0.25 \cdot 4 + 0.3 \cdot 5 + 0.18 \cdot 6 \\ &+ 0.05 \cdot 7 + 0.03 \cdot 8 + 0.02 \cdot 9 + 0.01 \cdot 10 = 4.84 < 5.24. \end{aligned}$$

This shows that *A* is a better marksman.

1.2. Properties of probability. Addition and multiplication of events. Incompatible and independent events

From the definition of probability adduced in the preceding section, it follows that the probability $p(A)$ of every event *A* is a real number in the range of 0 and 1:

$$0 \leq p(A) \leq 1.$$

Moreover, the probability may be 1, signifying that the event *A* is realized for

every outcome of the experiment under consideration, i.e., that A is the *certain* or *sure* event (thus, for example, the probability of drawing a white ball from an urn containing only white balls is 1). The probability may also be 0, implying that the event is not realized *for any* outcome of the experiment, i.e., it is *impossible* (the probability of drawing a black ball from an urn containing only white balls is 0).

Now, suppose that an experiment has only two mutually exclusive outcomes A and B . In such a case, B is called the *contrary* event of A and is denoted by \bar{A} (read as 'not- A '). If the event A is realized in m out of n equally probable outcomes of an experiment, then the event \bar{A} is realized in the remaining $n - m$ outcomes. Hence,

$$p(A) = m/n, \quad p(\bar{A}) = (n - m)/n = 1 - (m/n).$$

Consequently,

$$p(\bar{A}) = 1 - p(A).$$

Thus, the table of probabilities for an experiment having exactly two outcomes takes the simple form

A	\bar{A}
$p(A)$	$1 - p(A)$

Let us now consider two events A and A_1 such that the occurrence of A necessarily implies the occurrence of A_1 (for example, A is the appearance of a six in rolling a die and A_1 the appearance of a number divisible by 3). In such a case, obviously A_1 must occur in all those outcomes of the experiment in which the event A is realized. Hence, the probability of A_1 *cannot be less* than that of A . The situation in which the occurrence of A implies that of A_1 we write in symbols as $A \subset A_1$ (read as " A implies A_1 "). We have, thus, the following important property of probability:

$$\text{if } A \subset A_1, \text{ then } p(A) \leq p(A_1).$$

We consider the event, which consists of the occurrence of *at least one* of the two fixed events A and B . We call this event the *sum* of events A and B and denote it by $A + B$. For this, there are two basically distinct cases. If the events A and B are incompatible, i.e., it is impossible for both of them to occur simultaneously, then A occurs in any m_1 out of n equally probable outcomes of an experiment and B in the *different* m_2 outcomes. We have in this case

$$p(A) = \frac{m_1}{n}, \quad p(B) = \frac{m_2}{n} \text{ and } p(A + B) = \frac{m_1 + m_2}{n} = \frac{m_1}{n} + \frac{m_2}{n},$$

i.e.,

$$p(A + B) = p(A) + p(B)$$

(the *addition law of probabilities*). Thus, in the example considered on p. 3, the probability of drawing a *white* or a *black* ball, by virtue of this law, is,

$$\frac{1}{2} + \frac{3}{10} = \frac{4}{5}.$$

The addition law of probabilities formulated above may be generalized as follows. Suppose that we have k events A_1, A_2, \dots, A_k , *any two of which are pairwise incompatible*. We denote, by $A_1 + A_2 + \dots + A_k$, an event which consists in that at least one of these k events occurs. Then, obviously,

$$p(A_1 + A_2 + \dots + A_k) = p(A_1) + p(A_2) + \dots + p(A_k).$$

This more general result is sometimes also called the *addition law of probabilities*. In particular, if an experiment has k (and only k) distinct *mutually exclusive* outcomes, then the probabilities corresponding to it are given by the table:

A_1	A_2	\dots	A_k
$p(A_1)$	$p(A_2)$	\dots	$p(A_k)$

where the *numbers appearing in the lower row sum to 1*, i.e.,

$$p(A_1) + p(A_2) + \dots + p(A_k) = 1.$$

This stems from the fact that $p(A_1) + p(A_2) + \dots + p(A_k) = p(A_1 + A_2 + \dots + A_k)$ and that $A_1 + A_2 + \dots + A_k$ is a sure event (because any one outcome of the experiment is certain to be realized), so that

$$p(A_1 + A_2 + \dots + A_k) = 1.$$

Let us now assume that the events A and B may be *compatible*, i.e., can be realized simultaneously. In this case, it is, however, impossible to assert that $p(A + B) = p(A) + p(B)$. Indeed, suppose that A occurs in m_1 and B in m_2 of n equally probable outcomes of the experiment. The event $A + B$ is realized if the outcome that takes place is either the one from the first m_1 or the one from the second m_2 ; however, since these two sets of m_1 and m_2 outcomes may have common events, the possibility is that the total number of outcomes in two sets may be less than $m_1 + m_2$. Thus, in the general case, all that we can assert is that *the probability of the sum of two events can never exceed the sum of their*

probabilities:

$$p(A + B) \leq p(A) + p(B)$$

(but $p(A + B) \geq p(A)$ and $p(A + B) \geq p(B)$, since $A \subset A + B$ and $B \subset A + B$ by the very definition of the sum of events). Similarly, for every k arbitrary events (not necessarily mutually exclusive), we have

$$p(A_1 + A_2 + \dots + A_k) \leq p(A_1) + p(A_2) + \dots + p(A_k).$$

The inequality $p(A + B) \leq p(A) + p(B)$ can be made slightly more precise. We define the *product* of two events A and B as an event wherein *both* the events are realized simultaneously and denote it by AB . Let us consider m_1 (correspondingly m_2) equally probable outcomes of an experiment in which the event A (correspondingly B) occurs; we assume that there occur precisely l outcomes contained both in the m_1 outcomes favourable to the occurrence of A and m_2 outcomes favourable to the occurrence of B . It is obvious that both the events A and B are simultaneously realized if and only if one of these l outcomes occurs. Hence $p(AB) = l/n$. On the other hand, if exactly l outcomes are contained both in the m_1 outcomes favourable to the occurrence of A and the m_2 outcomes favourable to the occurrence of B , then in all we have $m_1 + m_2 - l$ outcomes (since the sum $m_1 + m_2$ contains l outcomes which are thus counted twice). Therefore,

$$p(A + B) = \frac{m_1 + m_2 - l}{n} = \frac{m_1}{n} + \frac{m_2}{n} - \frac{l}{n},$$

and, consequently,

$$p(A + B) = p(A) + p(B) - p(AB).$$

It is seen that the problem of determining the probability of the sum $A + B$ of events A and B reduces to the evaluation of the probability of the *product* AB of these events. The latter problem is not quite simple in a general case and it will be considered in the next section. However, there is a particular case in which the evaluation of the probability of event AB does not present any difficulty. This is the case in which A and B are *independent* events, i.e., the case in which the result of an experiment with which the occurrence or nonoccurrence of the event A is associated is in no way influenced by the conditions of an experiment result the event B is connected. Thus, for instance, the events involved in drawing a black ball from different urns containing black and white balls are independent, but two successive draws of a black ball from *one* urn (without replacement of the ball drawn) are not independent events (since the result of the first draw alters the number of balls left in the urn and, hence, is reflected in the conditions of the second experiment).

Suppose that the event A occurs in m_1 out of n_1 equally probable outcomes of the first experiment and, independently of this, the event B occurs in m_2 out of n_2 equally probable outcomes of the second experiment. Then, the probability of A is m_1/n_1 and that of B is m_2/n_2 . We now consider a compound experiment consisting of both the experiments under discussion. It is obvious that this compound experiment can have $n_1 n_2$ distinct equally probable outcomes, since to each of n_1 distinct outcomes of the first experiment we can associate distinct n_2 outcomes of the second experiment. Of these $n_1 n_2$ equally probable outcomes, $m_1 m_2$ equally probable outcomes are favourable to the occurrence of AB . These are obtained by combining the m_1 outcomes of the first experiment favourable to A with the m_2 outcomes of the second experiment favourable to B . The probability of the event AB is thus given by

$$\frac{m_1 m_2}{n_1 n_2} = \frac{m_1}{n_1} \cdot \frac{m_2}{n_2},$$

and, hence,

$$p(AB) = p(A) p(B)$$

(the multiplication law of probabilities).

This law can now be generalized as follows. Suppose that A_1, A_2, \dots, A_k are any k mutually independent events, i.e. the conditions of experiments to which the outcome of a particular event is related depend in no way upon the occurrence or nonoccurrence of the remaining events. In such a case,

$$p(A_1 A_2 \dots A_k) = p(A_1) p(A_2) \dots p(A_k).$$

The proof of this relation is exactly the same as the derivation of the formula $p(AB) = p(A) p(B)$, which forms its particular case.

If the events A and B are *not independent*, then the multiplication law $p(AB) = p(A) p(B)$ is not guaranteed. For example, if $B \subset A$ (say, A is the appearance of an even number in a die roll and B that of a two), then the event AB coincides with B and, consequently, $p(AB) = p(B)$. In fact, we can only assert that $p(AB) \leq p(A)$ and $p(AB) \leq p(B)$ (since from the definition of the product of events it follows that $AB \subset B$ and $AB \subset A$). The question concerning the evaluation of the product of two arbitrary events will be dealt with in more detail in the next section.

A few problems now follow to make obvious the applications of simple properties of probability we have deduced.

Problem 5. *A coin is flipped 2 times. What is the probability that a head occurs on both the flips?*

We seek here the probability of the events AB where A is the occurrence of a head on the first flip and B is the occurrence of the same face, that is, a head on

the second flip. The events A and B are obviously independent. Hence,

$$p(AB) = p(A) p(B) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

(see Problem 1 on p. 3).

Problem 6. *We select at random a positive integer not exceeding 1000. What is the probability that the selected positive integer can be expressed as a power of another integer (with exponent greater than unity)?*

The term 'at random' in the formulation of this problem implies that we regard the appearance of any number between 1 and 1000 to be equally probable. Furthermore, since

$$\begin{aligned} 2^9 < 1000 < 2^{10}, 3^6 < 1000 < 3^7, 5^4 < 1000 < 5^5, 6^3 < 1000 < 6^4, \\ 7^3 < 1000 < 7^4, 10^3 < 1000 < 10^4, 11^2 < 1000 < 11^3, \\ 12^2 < 1000 < 12^3, \dots, 31^2 < 1000 < 31^3, 32^2 > 1000, \end{aligned}$$

the probability that the selected integer will be a power of 2 is 8/1000 (among 1000 integers between 1 and 1000, there are 8 which occur as power of two: $2^2 = 4, 2^3 = 8, 2^4 = 16, 2^5 = 32, 2^6 = 64, 2^7 = 128, 2^8 = 256, 2^9 = 512$); in exactly the same way, the probabilities that the selected integer will equal the number 3, 5, 6, 7, 10, 11, 12, 13, 14, 15, 17, 18, 19, 20, 21, 22, 23, 24, 26, 28, 29, 30 and 31 raised to the integer power greater than 1 are correspondingly given by

$$\frac{5}{1000}, \frac{3}{1000}, \frac{2}{1000}, \frac{2}{1000}, \frac{2}{1000}, \frac{1}{1000}, \frac{1}{1000}, \dots, \frac{1}{1000}$$

(if the numbers raised to a power were 4, 8, 9, 16, 25 and 27, then they constitute simultaneously a smaller number raised to a greater power; hence, these cases have been excluded). Since, all the corresponding events are pairwise incompatible, the desired probability is given by

$$\begin{aligned} \frac{8}{1000} + \frac{5}{1000} + \frac{3}{1000} + \frac{2}{1000} + \frac{2}{1000} + \frac{2}{1000} \\ + \underbrace{\frac{1}{1000} + \frac{1}{1000} + \dots + \frac{1}{1000}}_{18 \text{ times}} = \frac{40}{1000} = \frac{1}{25}. \end{aligned}$$

Problem 7. *In a 52-card deck, cards of one of the four suits are specified as trumps. What is the probability that a card selected at random is either an ace or a trump?*

Suppose that the event A (correspondingly B) is that the card drawn be an ace (correspondingly, trump); then the event AB is that this card be an ace of trumps

and $p(A) = \frac{1}{13}$ (each suit of the deck contains 13 cards : two, three, . . . , ace), $p(B) = \frac{1}{4}$, $p(AB) = \frac{1}{52}$. Hence, the desired probability is given by

$$p(A + B) = p(A) + p(B) - p(AB) = \frac{1}{13} + \frac{1}{4} - \frac{1}{52} = \frac{4}{13}.$$

Problem 8. *Six hunters saw a fox and simultaneously shot at it. It is assumed that each of them, as a rule, hits the fox at a given distance and kills it in one out of three chances. What is the probability that the fox is killed?*

Suppose that A_1, A_2, \dots, A_6 denote the events that the fox is killed by the 1st, 2nd, . . . , 6th hunter. In the hypothesis of the problem, it is indicated that

$$p(A_1) = p(A_2) = \dots = p(A_6) = \frac{1}{3};$$

it is required to find $p(S)$, where $S = A_1 + A_2 + \dots + A_6$. The events A_1, A_2, \dots, A_6 are obviously independent; this enables us to solve this problem by multiple use of the formula

$$p(A + B) = p(A) + p(B) - p(AB) = p(A) + p(B) - p(A)p(B)$$

(see the discussion below in small type). However, such a solution is not simple since the formula expressing the probability of the sum of several (compatible) events is fairly complicated.

The following version of the solution of this problem is more convenient. Let us first determine the probability $p(\bar{S})$ that the fox escapes. A miss by the 1st, 2nd, . . . , 6th hunter is naturally denoted by $\bar{A}_1, \bar{A}_2, \dots, \bar{A}_6$; by the formula $p(\bar{A}) = 1 - p(A)$, we have

$$p(\bar{A}_1) = p(\bar{A}_2) = \dots = p(\bar{A}_6) = \frac{2}{3}.$$

In order for the fox to survive, it is necessary that *all* the hunters miss the target, i.e., the problem here relates to the probability of the product events $\bar{A}_1, \bar{A}_2, \dots, \bar{A}_6$, where all the events $\bar{A}_1, \bar{A}_2, \dots, \bar{A}_6$ are mutually independent. Thus,

$$p(\bar{S}) = p(\bar{A}_1 \bar{A}_2 \dots \bar{A}_6) = p(\bar{A}_1) \cdot \dots \cdot p(\bar{A}_6) = \frac{2}{3} \cdot \frac{2}{3} \cdot \dots \cdot \frac{2}{3} = \frac{2^6}{3^6} = \frac{64}{729},$$

and recalling the formula $p(\bar{A}) = 1 - p(A)$,

$$p(S) = 1 - \frac{64}{729} = \frac{665}{729}.$$

The formula

$$p(A + B) = p(A) + p(B) - p(AB)$$

can be extended also to the case of determining the probability of the sum of an arbitrary number k of (possibly, compatible) events A_1, A_2, \dots, A_k . We have

$$p(A_1 + A_2 + A_3) = p\{(A_1 + A_2) + A_3\} = p(A_1 + A_2) + p(A_3) - p\{(A_1 + A_2) A_3\}.$$

Here

$$p(A_1 + A_2) = p(A_1) + p(A_2) - p(A_1 A_2).$$

Let us now explain the meaning of the more complicated expression $p\{(A_1 + A_2) A_3\}$. By the definition of the sum and product of events, the event $(A_1 + A_2) A_3$ consists of the occurrence of *at least one* of the events A_1 and A_2 and simultaneously the event A_3 also occurs. But this means that at least one of the events $A_1 A_3$ and $A_2 A_3$ occurs. Thus, we have

$$(A_1 + A_2) A_3 = A_1 A_3 + A_2 A_3,$$

and, consequently,

$$p\{(A_1 + A_2) A_3\} = p(A_1 A_3 + A_2 A_3) = p(A_1 A_3) + p(A_2 A_3) - p\{(A_1 A_3)(A_2 A_3)\}.$$

Furthermore, the event $(A_1 A_3)(A_2 A_3)$ consists of the simultaneous occurrence of both the events $A_1 A_3$ (i.e., A_1 as well as A_3) and $A_2 A_3$ (A_2 as well as A_3). In other words, the event $(A_1 A_3)(A_2 A_3)$ consists of the simultaneous occurrence of *three* events A_1, A_2 and A_3 , i.e., it does not differ from the event $A_1 A_2 A_3$.

We thus finally obtain

$$p(A_1 + A_2 + A_3) = p(A_1) + p(A_2) - p(A_1 A_2) + p(A_3) - p(A_1 A_3) - p(A_2 A_3) + p(A_1 A_2 A_3),$$

or, in a different order,

$$p(A_1 + A_2 + A_3) = p(A_1) + p(A_2) + p(A_3) - p(A_1 A_2) - p(A_1 A_3) - p(A_2 A_3) + p(A_1 A_2 A_3).$$

Proceeding on lines similar to this, for arbitrary k , we have

$$\begin{aligned} p(A_1 + A_2 + \dots + A_k) &= p(A_1) + p(A_2) + \dots + p(A_k) - p(A_1 A_2) - p(A_1 A_3) \\ &\quad - \dots - p(A_{k-1} A_k) + p(A_1 A_2 A_3) + p(A_1 A_2 A_4) \\ &\quad + \dots + p(A_{k-2} A_{k-1} A_k) - p(A_1 A_2 A_3 A_4) \\ &\quad - \dots + (-1)^{k-1} p(A_1 A_2 \dots A_k). \end{aligned}$$

This formula can be easily proved by induction, following a procedure similar to the one demonstrated above for a case when $k = 3$.

Let us now solve Problem 8 with the aid of the formula deduced. For $k = 6$,

$$\begin{aligned} p(A_1 + A_2 + \dots + A_6) &= p(A_1) + p(A_2) + \dots + p(A_6) - p(A_1 A_2) - p(A_1 A_3) \\ &\quad - \dots - p(A_5 A_6) + p(A_1 A_2 A_3) + p(A_1 A_2 A_4) + \dots \\ &\quad + p(A_4 A_5 A_6) - \dots - p(A_1 A_2 A_3 A_4 A_5 A_6). \end{aligned}$$

But (since all the events A_1, A_2, \dots, A_k are mutually independent),

$$p(A_1) = p(A_2) = \dots = p(A_6) = \frac{1}{3}$$

$$p(A_1 A_2) = p(A_1 A_3) = \dots = p(A_5 A_6) = p(A_1) p(A_2) = \left(\frac{1}{3}\right)^2,$$

$$p(A_1 A_2 A_3) = \dots = p(A_4 A_5 A_6) = p(A_1) p(A_2) p(A_3) = \left(\frac{1}{3}\right)^3, \dots,$$

$$p(A_1 A_2 \dots A_6) = p(A_1) p(A_2) \dots p(A_6) = \left(\frac{1}{3}\right)^6,$$

hence we get

$$\begin{aligned} p(A_1 + A_2 + \dots + A_6) &= 6 \cdot \frac{1}{3} - C\left(\frac{6}{2}\right)\left(\frac{1}{3}\right)^2 + C\left(\frac{6}{3}\right)\left(\frac{1}{3}\right)^3 \\ &\quad - C\left(\frac{6}{4}\right)\left(\frac{1}{3}\right)^4 + C\left(\frac{6}{5}\right)\left(\frac{1}{3}\right)^5 - C\left(\frac{6}{6}\right)\left(\frac{1}{3}\right)^6 \\ &= 1 - \left(1 - \frac{1}{3}\right)^6 = 1 - \left(\frac{2}{3}\right)^6 = \frac{665}{729}, \end{aligned}$$

i.e. the same result as above.

Other examples of the applications of this general formula can be found in [40].

We now turn to the concepts of the 'sum' and 'product' of *random variables*, which will be put to good use in the sequel as well. We illustrate the former by the following example.

Problem 9. *Two different lathes are installed in a workshop, manufacturing identical parts. It is known from experience that the first (older) lathe may produce up to three defective parts in a day, the probability of the number of defective parts being given here by:*

Number of defective parts (per day)	0	1	2	3
Probability	0.3	0.4	0.2	0.1

The second (new) lathe produces not more than one defective part per day. The probability that at most one of the parts manufactured in a day is found defective is in all equal to 0.1:

Number of defective parts (per day)	0	1
Probability	0.9	0.1

The question is: What is the average number of defective parts manufactured per day in the workshop.

In this problem, we simultaneously consider two random variables α and β . The former variable α takes the values a_0, a_1, a_2 , and a_3 (precisely, 0, 1, 2 and 3) with the probabilities p_0, p_1, p_2 and p_3 (which are equal in our case to 0.3, 0.4, 0.2 and 0.1; obviously $p_0 + p_1 + p_2 + p_3 = 1$, as it must be). The latter variable β takes only two values b_0 and b_1 (namely, 0 and 1) with the probabilities q_0 and q_1 (in this case 0.9 and 0.1; clearly, $q_0 + q_1 = 1$, as it must be). The mean values of these random variables (i.e., α and β) represent the average number of defective parts produced in a day, respectively, by the first and second lathes; they are correspondingly given by

$$\begin{aligned} \text{m.v. } \alpha &= p_0 a_0 + p_1 a_1 + p_2 a_2 + p_3 a_3 = 0.3 \cdot 0 + 0.4 \cdot 1 + 0.2 \cdot 2 \\ &\quad + 0.1 \cdot 3 = 1.1, \end{aligned}$$

and

$$\text{m.v. } \beta = q_0 b_0 + q_1 b_1 = 0.9 \cdot 0 + 0.1 \cdot 1 = 0.1.$$

We are, however, interested in the random variable $\alpha + \beta$, representing the number of defective parts produced in a day by both the lathes. This variable can take the values

$$a_0 + b_0, a_0 + b_1; a_1 + b_0, a_1 + b_1; a_2 + b_0, a_2 + b_1; a_3 + b_0 \text{ and } a_3 + b_1$$

(in the present case, the values 0, 1, 2, 3 and 4). We shall assume (for a while !) that the random variables α and β are *independent*, i.e., we assume that the random variable α takes the values 0, 1, 2 and 3 with the probabilities p_0, p_1, p_2 and p_3 (i.e., 0.3, 0.4, 0.2 and 0.1) irrespective of the value which is taken by the variable β (for the same day). Then, the events $\alpha = a_i$ ($i = 0, 1, 2$ or 3) and $\beta = b_j$ ($j = 0$ or 1) are also *independent*, and hence,

$$p(\alpha = a_i \text{ and } \beta = b_j) = p(\alpha = a_i) \cdot p(\beta = b_j) = p_i q_j.$$

This yields the following (detailed) probability table of the random variable $\alpha + \beta$:

Values	$a_0 + b_0$ (= 0)	$a_0 + b_1$ (= 1)	$a_1 + b_0$ (= 1)	$a_1 + b_1$ (= 2)	$a_2 + b_0$ (= 2)	$a_2 + b_1$ (= 3)	$a_3 + b_0$ (= 3)	$a_3 + b_1$ (= 4)
Probabilities	$p_0 q_0$ (= 0.27)	$p_0 q_1$ (= 0.03)	$p_1 q_0$ (= 0.36)	$p_1 q_1$ (= 0.04)	$p_2 q_0$ (= 0.18)	$p_2 q_1$ (= 0.02)	$p_3 q_0$ (= 0.09)	$p_3 q_1$ (= 0.01)

Hence

$$\begin{aligned} \text{m.v. } (\alpha + \beta) &= p_0 q_0 (a_0 + b_0) + p_0 q_1 (a_0 + b_1) + p_1 q_0 (a_1 + b_0) \\ &\quad + p_1 q_1 (a_1 + b_1) + p_2 q_0 (a_2 + b_0) + p_2 q_1 (a_2 + b_1) \\ &\quad + p_3 q_0 (a_3 + b_0) + p_3 q_1 (a_3 + b_1) \\ &= a_0 (p_0 q_0 + p_0 q_1) + a_1 (p_1 q_0 + p_1 q_1) + a_2 (p_2 q_0 + p_2 q_1) \\ &\quad + a_3 (p_3 q_0 + p_3 q_1) + b_0 (p_0 q_0 + p_1 q_0 + p_2 q_0 + p_3 q_0) \\ &\quad + b_1 (p_0 q_1 + p_1 q_1 + p_2 q_1 + p_3 q_1) \\ &= a_0 p_0 (q_0 + q_1) + a_1 p_1 (q_0 + q_1) + a_2 p_2 (q_0 + q_1) \\ &\quad + a_3 p_3 (q_0 + q_1) + b_0 q_0 (p_0 + p_1 + p_2 + p_3) \\ &\quad + b_1 q_1 (p_0 + p_1 + p_2 + p_3) \\ &= (a_0 p_0 + a_1 p_1 + a_2 p_2 + a_3 p_3) + (b_0 q_0 + b_1 q_1) \\ &= \text{m.v. } \alpha + \text{m.v. } \beta = 1.2 \text{ (defective parts per day).} \end{aligned}$$

It is thus seen that the *mean value of the sum of two random variables is the sum of their mean values*.

However, it is worthwhile to note that the last conclusion obtained via sufficiently tedious algebraic transformations is in fact quite elementary. Suppose that on a specific day, say the first day, the first lathe produces $a^{(1)}$ defective parts (where $a^{(1)}$ equals 0, 1, 2 or 3) and the second lathe produces $b^{(1)}$ defective parts

(where $b^{(1)}$ equals 0 or 1). Let it be further assumed that on the second, third, . . . , n th day, the first lathe produces $a^{(2)}, a^{(3)}, \dots, a^{(n)}$ defective parts and the second lathe produces $b^{(2)}, b^{(3)}, \dots, b^{(n)}$ such parts. Then, the total number of defective parts produced on the first, second, third, . . . , n th day is given by

$$a^{(1)} + b^{(1)}, a^{(2)} + b^{(2)}, a^{(3)} + b^{(3)}, \dots, a^{(n)} + b^{(n)},$$

and the *mean number* of defective parts produced per day is given by

$$\begin{aligned} & \frac{(a^{(1)} + b^{(1)}) + (a^{(2)} + b^{(2)}) + (a^{(3)} + b^{(3)}) + \dots + (a^{(n)} + b^{(n)})}{n} \\ &= \frac{a^{(1)} + a^{(2)} + a^{(3)} + \dots + a^{(n)}}{n} + \frac{b^{(1)} + b^{(2)} + b^{(3)} + \dots + b^{(n)}}{n}. \end{aligned}$$

But, for large n , the value

$$\frac{(a^{(1)} + b^{(1)}) + (a^{(2)} + b^{(2)}) + (a^{(3)} + b^{(3)}) + \dots + (a^{(n)} + b^{(n)})}{n}$$

will be very close to the m.v. $(\alpha + \beta)$, and the values

$$\frac{a^{(1)} + a^{(2)} + a^{(3)} + \dots + a^{(n)}}{n},$$

and

$$\frac{b^{(1)} + b^{(2)} + b^{(3)} + \dots + b^{(n)}}{n}$$

to m.v. α and m.v. β . This fact obviously implies that

$$\text{m.v. } (\alpha + \beta) = \text{m.v. } \alpha + \text{m.v. } \beta.$$

It is noteworthy that the conclusion established by the preceding simple reasoning is more general than that proved algebraically earlier? In fact, in this reasoning we did not have to rely on the *independence* of the variables α and β (which, as a matter of fact, is not tenable rather often in practice, because the operations of both lathes can be affected by certain common factors such as, e.g., use of the same raw material by both lathes). Therefore, it is impossible to assert in a general case that

$$p(\alpha = a_i \text{ and } \beta = b_j) = p(\alpha = a_i) \cdot p(\beta = b_j) = p_i q_j.$$

Hence, in place of the values $p_0 q_0$, $p_0 q_1$ and so on, the second row of the probability table of the random variable $\alpha + \beta$ will contain certain probabilities p_{00} (the probability that $\alpha = a_0$ and $\beta = b_0$), p_{01} (the probability that $\alpha = a_0$ and

$\beta = b_1$) and so on whose numerical values depend on the relationship between the variables α and β , usually unknown to us in all details.

This situation, however, has almost no impact on the calculations adduced earlier. In fact, we now have

$$\begin{aligned} \text{m.v. } (\alpha + \beta) &= p_{00}(a_0 + b_0) + p_{01}(a_0 + b_1) + p_{10}(a_1 + b_0) + p_{11}(a_1 + b_1) \\ &\quad + p_{20}(a_2 + b_0) + p_{21}(a_2 + b_1) + p_{30}(a_3 + b_0) + p_{31}(a_3 + b_1) \\ &= a_0(p_{00} + p_{01}) + a_1(p_{10} + p_{11}) + a_2(p_{20} + p_{21}) \\ &\quad + a_3(p_{30} + p_{31}) + b_0(p_{00} + p_{10} + p_{20} + p_{30}) \\ &\quad + b_1(p_{01} + p_{11} + p_{21} + p_{31}). \end{aligned}$$

But

$$\begin{aligned} p_{00} + p_{01} &= p(\alpha = a_0 \text{ and } \beta = b_0) + p(\alpha = a_0 \text{ and } \beta = b_1) \\ &= p(\alpha = a_0 \text{ and } \beta = b_0 \text{ or } b_1). \end{aligned}$$

However, b_0 and b_1 represent *all* possible values of the random variable β and hence $p(\alpha = a_0 \text{ and } \beta = b_0 \text{ or } b_1)$ is nothing more than simply $p(\alpha = a_0) = p_0$! In precisely the same way, it is established that

$$p_{10} + p_{11} = p_1, \quad p_{20} + p_{21} = p_2, \quad p_{30} + p_{31} = p_3.$$

Furthermore,

$$\begin{aligned} p_{00} + p_{10} + p_{20} + p_{30} &= p(\alpha = a_0 \text{ and } \beta = b_0) + p(\alpha = a_1 \text{ and } \beta = b_0) \\ &\quad + p(\alpha = a_2 \text{ and } \beta = b_0) + p(\alpha = a_3 \text{ and } \beta = b_0) \\ &= p(\alpha = a_0 \text{ or } a_1 \text{ or } a_2 \text{ or } a_3 \text{ and } \beta = b_0) \\ &= p(\beta = b_0) = q_0, \end{aligned}$$

and similarly

$$p_{01} + p_{11} + p_{21} + p_{31} = q_1.$$

Thus, as before, in this case we have

$$\begin{aligned} \text{m.v. } (\alpha + \beta) &= (a_0p_0 + a_1p_1 + a_2p_2 + a_3p_3) + (b_0q_0 + b_1q_1) \\ &= \text{m.v. } \alpha + \text{m.v. } \beta. \end{aligned}$$

The results obtained can, of course, be extended to *any* number of random variables that likewise satisfy the condition that the *mean value of their sum be equal to the sum of their mean values*.

We now return to the notion of the *product* of two random variables and put this to work in the following example.

Problem 10. *Every year a farmer sends a_0, a_1, a_2 or a_3 calves to a market and the probability (frequency) of a specific number of calves being sold is given by*

Number of calves	a_0	a_1	a_2	a_3
Probability	p_0	p_1	p_2	p_3

(where, of course, $p_0 + p_1 + p_2 + p_3 = 1$). On the other hand, the price fetched by a calf in different years may equal to either b_0 or b_1 , the probability of these prices being, respectively, equal to q_0 and q_1 ($= 1 - q_0$):

Price of calf	b_0	b_1
Probability	q_0	q_1

Find the farmer's mean annual receipt from the sale of calves.

Here, we are again concerned with the two random variables α and β . Retaining an analogy with Problem 9, we adhere to the same symbols as above and denote the possible values of these variables and the probabilities of these values by $a_0, a_1, a_2, a_3; b_0, b_1$ and $p_0, p_1, p_2, p_3; q_0, q_1$. Now, we are interested in the product $\alpha\beta$ of these two variables (the product of the number of calves sold and the price fetched by a calf), which can take 8 values $a_0b_0, a_0b_1; a_1b_0, a_1b_1; a_2b_0, a_2b_1; a_3b_0, a_3b_1$. In addition, if we consider α and β to be independent, then the probability table of the variable $\alpha\beta$ has the form

Values	a_0b_0	a_0b_1	a_1b_0	a_1b_1	a_2b_0	a_2b_1	a_3b_0	a_3b_1
Probability	p_0q_0	p_0q_1	p_1q_0	p_1q_1	p_2q_0	p_2q_1	p_3q_0	p_3q_1

Hence, the mean value of $\alpha\beta$ in this case is given by

$$\begin{aligned}
 \text{m.v. } (\alpha\beta) &= p_0q_0a_0b_0 + p_0q_1a_0b_1 + p_1q_0a_1b_0 + p_1q_1a_1b_1 + p_2q_0a_2b_0 \\
 &\quad + p_2q_1a_2b_1 + p_3q_0a_3b_0 + p_3q_1a_3b_1 \\
 &= p_0a_0(q_0b_0 + q_1b_1) + p_1a_1(q_0b_0 + q_1b_1) \\
 &\quad + p_2a_2(q_0b_0 + q_1b_1) + p_3a_3(q_0b_0 + q_1b_1) \\
 &= (p_0a_0 + p_1a_1 + p_2a_2 + p_3a_3)(q_0b_0 + q_1b_1) \\
 &= (\text{m.v. } \alpha) \cdot (\text{m.v. } \beta).
 \end{aligned}$$

It is thus seen that, for independent random variables α and β , the mean value of their product always equals the product of the mean values of these variables. The same principle also holds for any number of mutually independent random variables; here also the mean value of the product of all variables equals the product of the mean values of all the factor variables.

It may be remarked that in contrast to the case of the sum of random variables, in the case of the product the independence of factor variables is an essential condition, without which the results stated above can be found to be false. To illustrate this, it suffices to consider the case in which $\alpha_1 = \alpha_2 = \alpha$, where α is characterized by the following probability table:

Values of the variable α	+1	-1
Probability	0.5	0.5

In this case it is obvious that

$$\text{m.v. } \alpha_1 = \text{m.v. } \alpha_2 = 0.5(+1) + 0.5(-1) = 0,$$

so that

$$(\text{m.v. } \alpha_1) \times (\text{m.v. } \alpha_2) = 0 \times 0 = 0.$$

It is also evident that the variable $\alpha_1 \times \alpha_2 = \alpha^2$ is always equal to $+1$ (since $(+1)^2 = (-1)^2 = +1$), so that

$$\text{m.v. } (\alpha_1 \alpha_2) = 1 > 0 = (\text{m.v. } \alpha_1) \times (\text{m.v. } \alpha_2).$$

The inequality

$$\text{m.v. } (\alpha^2) > (\text{m.v. } \alpha)^2$$

established by this example will be revisited in Sec. 4 of this chapter.

1.3. Conditional probability

Two events A and B are called *independent*, if the result of the experiment to which A is related has no influence on the realization of the experiment with which B is associated. However, this situation does not always hold at all. An example substantiating this statement has been given earlier and will be reiterated here in detail. Suppose that event A consists of drawing a black ball from an urn containing m black and $n - m$ white balls and event B of drawing a black ball from the same urn *after* one ball is drawn. It is obvious that, if the first ball drawn is black, i.e., if A occurs, then after the first draw, $m - 1$ black and $n - m$ white balls are left in the urn and, hence, the probability of event B is $(m - 1)/(n - 1)$. If, however, the first ball drawn is white (namely, the event \bar{A} occurs), then m black and $n - m - 1$ white balls are left in the urn, and the desired probability equals $m/(n - 1)$. The probability of event B thus varies according as A is realized, or not, i.e., here the probability of event B can take two different values $[(m - 1)/(n - 1)]$ and $[m/(n - 1)]$, for which it is necessary also to prescribe separate notations.

The probability of event B in the case when it is known that event A has occurred is called the *conditional probability of B on the hypothesis that A has materialized* and is written as $p_A(B)$. Thus, in our case $p_A(B) = (m - 1)/(n - 1)$. Similarly, we define the conditional probability $p_{\bar{A}}(B)$ of B under the assumption that \bar{A} has occurred (i.e., under the assumption that A has not occurred); in our case $p_{\bar{A}}(B) = [m/(n - 1)]$.

It is also obvious that the conditional probability $p_A(B)$ of any event B under the assumption that A has occurred can be obviously either less or greater than the unconditional probability $p(B)$ of this event (i.e., the probability of B when

nothing is known about the result of the experiment involving A). Thus, in the example considered above, it is clear that $p(B) = m/n$, since it is possible to anticipate beforehand with equal probability that in the second draw any of the n balls contained in the urn will be drawn, and out of these n balls precisely m are black. Thus, here

$$p_A(B) = \frac{m-1}{n-1} < \frac{m}{n} = p(B) \quad \text{and} \quad p_{\bar{A}}(B) = \frac{m}{n-1} > \frac{m}{n} = p(B).$$

If A and B are independent events, then, obviously $p_A(B) = p(B)$. The last specification can be regarded as a *precise mathematical definition* of the notion of independence of events and it enables us to verify for any pair of events A and B whether they are independent or not (see in this context, the specific example given at the end of this section in small type).

Conditional probabilities can be calculated quite similarly the way we computed unconditional probabilities in Sec. 1. Suppose that event A has N equally probable outcomes of an experiment favourable to it, which permit us to determine the occurrence or nonoccurrence of A and also of a certain other event B . Out of these N outcomes, let M be favourable also to B so that the remaining $N - M$ are not favourable to B . In this case

$$p_A(B) = \frac{M}{N} \quad \left(\text{and } p_A(\bar{B}) = \frac{N-M}{N} \right).$$

Thus, for instance, in the example examined above, the experiment consisting of the successive draw of two balls from an urn with n balls, has $n(n-1)$ equally probable outcomes (in the first draw, any of n existing balls may be drawn and in the second, one of the remaining $n-1$). Out of these $n(n-1)$ outcomes there are $N = m(n-1)$ outcomes favourable to A (the first draw resulting in one of m black balls followed by any of the remaining $n-1$ balls); moreover, of these $m(n-1)$ outcomes favourable to A , those favourable to B are $M = m(m-1)$ (the first draw resulting in any of m black balls and the succeeding one in any of the remaining $m-1$ black balls). Consequently,

$$p_A(B) = \frac{M}{N} = \frac{m(m-1)}{m(n-1)} = \frac{m-1}{n-1}.$$

Let us now call K the total number of equally probable outcomes of an experiment with which is associated the occurrence of two events A and B . Since out of these K outcomes M are favourable to the occurrence of both A and B , the probability of the event AB , i.e., of the occurrence of both A and B , equals M/K . However, $M/K = (N/K) \times (M/N)$, but $M/N = p_A(B)$ and $N/K = p(A)$ (because out of K equally probable outcomes, N are favourable to A).

Consequently, we have

$$p(AB) = p(A) p_A(B).$$

This is also the general rule for the determination of the probability of the product AB of two events, usually called the *multiplication law of probabilities* (the multiplication law of Sec. 2 being a particular case). Thus, in order to find $p(AB)$, it is necessary to know the conditional probability $p_A(B)$, which characterizes the relationship existing between the events A and B . Therefore, the probability of AB is in general not determined by both the probabilities $p(A)$ and $p(B)$. Only in the case in which the probability of B is not affected as a result of the occurrence or nonoccurrence of event A , i.e., in which A and B are *independent*, we have $p_A(B) = p(B)$ and $p(AB) = p(A) p(B)$, the conclusions that we obtained above.

From the definition of conditional probability, we immediately deduce the following properties:

- (a) $0 \leq p_A(B) \leq 1$; $p_A(B) = 1$, if $A \subset B$ (in particular, if B is the certain event); $p_A(B) = 0$, if A and B are incompatible (in particular, if B is the impossible event);
- (b) if $B \subset B_1$, then $p_A(B) \leq p_A(B_1)$;
- (c) if B and C are incompatible, then $p_A(B + C) = p_A(B) + p_A(C)$; if B_1, B_2, \dots, B_k are pairwise incompatible, then

$$p_A(B_1 + B_2 + \dots + B_k) = p_A(B_1) + p_A(B_2) + \dots + p_A(B_k);$$
- (d) $p_A(\bar{B}) = 1 - p_A(B)$.

The proof of these properties is completely analogous to the proof deduced in Sec. 2 for these very properties for ordinary (unconditional) probabilities.

We further note that the formula $p(AB) = p(A) p_A(B)$ implies

$$p(A)p_A(B) = p(B)p_B(A), \quad \text{or} \quad \frac{p_B(A)}{p(A)} = \frac{p_A(B)}{p(B)}$$

(since it is obvious that the events AB and BA are identical ones). This implies, in particular, that knowing the probabilities $p(A)$ and $p(B)$ of two events A and B and the conditional probability $p_A(B)$ of B under the assumption that A occurs, it is possible also to determine the conditional probability $p_B(A)$:

$$p_B(A) = p_A(B) \times \frac{p(A)}{p(B)}.$$

Thus, in the urn example analyzed above, $p(A) = p(B) = m/n$ (the probabilities of drawing a black ball in the first and in the second draw both equal m/n); hence, $p_B(A) = p_A(B) = (m-1)/(n-1)$ (here $p_B(A)$ is the probability of

drawing a black ball in the first draw if it is known that the second draw results in a black ball).

We finally remark that since either one of the events A and \bar{A} necessarily occurs, the sum of events AB (i.e., ' B and A ') and $\bar{A}B$ (' B and \bar{A} ') coincides with the event B . But since

$$p(AB) = p(A) p_A(B), \quad p(\bar{A}B) = p(\bar{A}) p_{\bar{A}}(B),$$

and

$$p(AB + \bar{A}B) = p(AB) + p(\bar{A}B)$$

(events AB and $\bar{A}B$ are clearly incompatible, because A and \bar{A} are so), then

$$p(B) = p(A) p_A(B) + p(\bar{A}) p_{\bar{A}}(B).$$

Thus in the example under consideration,

$$p(A) = \frac{m}{n}, \quad p(\bar{A}) = \frac{n-m}{n}, \quad p_A(B) = \frac{m-1}{n-1}, \quad p_{\bar{A}}(B) = \frac{m}{n-1},$$

and

$$p(A) p_A(B) + p(\bar{A}) p_{\bar{A}}(B) = \frac{m}{n} \frac{m-1}{n-1} + \frac{n-m}{n} \frac{m}{n-1} = \frac{m}{n} = p(B).$$

Quite similarly, if any experiment α can have k (and only k) pairwise incompatible outcomes A_1, A_2, \dots, A_k , then every event B can be expressed as the sum of events $A_1B + A_2B + \dots + A_kB$. Hence

$$p(B) = p(A_1)p_{A_1}(B) + p(A_2)p_{A_2}(B) + \dots + p(A_k)p_{A_k}(B).$$

This equation is called the *equation of total probability*.

Problem 11. *There are three urns : urn 1 contains 2 white and 4 black balls, urn 2 contains 4 white and 2 black balls and urn 3 contains 3 white and 3 black balls.*

A ball is drawn at random from an urn (which urn is not known). What is the probability that the selected ball is from the first urn if it turns out to be (a) white, (b) black ?

Suppose that event A (resp. event \bar{A}) be that the selected ball is white (resp. black). Moreover, let event B be that the ball is removed from the first urn. Our experiment of drawing a single ball can have $3 \times 6 = 18$ outcomes (according to the total number of balls in all three of the urns) which we regarded to be equally probable (in other words, the drawing of a ball from any of the urns is considered to be equally probable). Of these 18 outcomes, 9 are favourable to

A and of the last 9 outcomes 2 are favourable to B . Of these 18 outcomes, 9 are favourable to \bar{A} , too, but of these 9 outcomes 4 are favourable to B . Thus, we have

$$p_A(B) = \frac{2}{9} \text{ and } p_{\bar{A}}(B) = \frac{4}{9}.$$

Problem 12. *A word, 'papagay' is formed by letters of an alphabetic section. Then, cards with the letters are well mixed and any four of them are drawn one after another in succession and arranged in a row. What is the probability of obtaining the word 'papa' by this procedure ?*

Suppose that events A, B, C and D be, respectively, that the first letter drawn is 'p'; the second 'a'; the third 'p' and the fourth 'a'; then the event in whose probability we are interested can be written as $ABCD$. Further, by applying consecutively a few times the formula for the probability of the product of two events, we have

$$p(A) = \frac{2}{7},$$

$$p(AB) = p(A) p_A(B) = \frac{2}{7} \times \frac{3}{6} = \frac{1}{7},$$

$$p(ABC) = p(AB) p_{AB}(C) = \frac{1}{7} \times \frac{1}{5} = \frac{1}{35},$$

and, finally,

$$p(ABCD) = p(ABC) p_{ABC}(D) = \frac{1}{35} \times \frac{2}{4} = \frac{1}{70}.$$

Problem 13. *We have 5 urns, of which two urns each contain 1 white and 5 black balls; one urn contains 2 white and 5 black balls, and, finally, each of the last two urns contains 3 white and 5 black balls. An urn is chosen at random and a ball is drawn at random from it. What is the probability that the selected ball is white ?*

We denote by A_1, A_2 and A_3 the events such that the ball is drawn from an urn containing, respectively, one, two, or three white balls. Then,

$$p(A_1) = \frac{2}{5}; \quad p(A_2) = \frac{1}{5}; \quad \text{and} \quad p(A_3) = \frac{2}{5}.$$

Further, if B is an event such that the selected ball is white, then by the equa-

tion of total probability, we have

$$\begin{aligned} p(B) &= p(A_1) \times p_{A_1}(B) + p(A_2) \times p_{A_2}(B) + p(A_3) \times p_{A_3}(B) \\ &= \frac{2}{5} \times \frac{1}{6} + \frac{1}{5} \times \frac{2}{7} + \frac{2}{5} \times \frac{3}{8} = \frac{23}{84}. \end{aligned}$$

We conclude with a simple example to demonstrate the application of the definition of independent random events given on p. 21. We consider a regular tetrahedron of homogeneous material, with the digits 1, 2 and 3 inscribed on its three faces and all these digits together on the fourth face (see Figure 2). Let A , B and C denote events such that the throw of the tetrahedron results in showing the face with digits 1, 2 and 3, respectively. It is, thus, obvious that

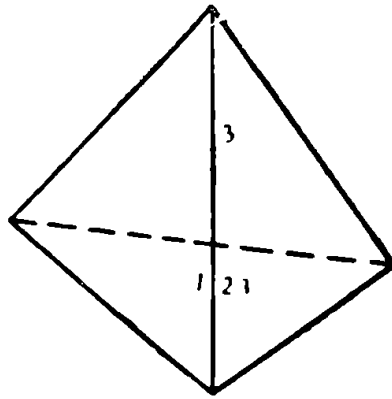


Fig. 2.

$p(A) = p(B) = p(C) = \frac{1}{2}$. Indeed, the tetrahedron may fall on any one of its faces with the same probability and each of the digits appears precisely on two of its four faces. If it is now known that event A has occurred, then it signifies the appearance of the face of tetrahedron, inscribed with either digit 1, or showing the three digits 1, 2 and 3. In addition, both the events B and C are realized in the latter but not in the former case. Consequently, $p_A(B) = p_A(C) = \frac{1}{2}$, so that

$$p_A(B) = p(B) \quad \text{and} \quad p_A(C) = p(C).$$

Hence both A and B , and A and C are *independent*, which yields also

$$p(AB) = p(A)p(B) = \frac{1}{4}, \quad p(AC) = p(A)p(C) = \frac{1}{4}$$

(see the multiplication law of probabilities for independent events on page 11). Similarly we can verify that the events B and C are also independent, i.e., here too we have $p_B(C) = p(C) = \frac{1}{2}$.

From the example adduced, we can also infer that pairwise independence of every pair of events among A , B and C does not *imply* the independence of all the three of them taken together, i.e., the validity of the equation

$$p(ABC) = p(A)p(B)p(C)$$

(cf. p. 11). It is, in fact, obvious that in our example the simultaneous occurrence of A and

B implies also the occurrence of C , so that here we have

$$p_{AB}(C) = 1 \quad \text{and} \quad p(ABC) = p(AB) p_{AB}(C) = \frac{1}{4} \times 1 = \frac{1}{4},$$

while

$$p(A) p(B) p(C) = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8}.$$

1.4. The variance of a random variable. Chebyshev's inequality and the law of large numbers

A very important characteristic of a random variable is, of course, its *average (mean) value*. With the aid of mean values, we can compare two random variables; thus, for example, between the two marksmen (see Problem 4, p. 7) the better shot is naturally the one who scores a higher mean number of points. There are, however, many problems where the knowledge of merely mean value of a random variable supplies very scanty information about the variable. We consider, for example, a cannon aimed to hit a target clamped into a vise at a distance a km from the cannon (Fig. 3). If we denote by α (km) the firing range of the shell, then the mean value of α , as a rule, equals a ; the deviation of the average value from a testifies to the presence of a *systematic* error in firing (systematic error in the flight of the shell beyond, or short of, the target), which can be eliminated by suitably changing the inclination of the barrel of the cannon. However, the absence of a systematic error does not at all guarantee high accuracy in firing. To evaluate accuracy, it is also necessary to know how close the shells come to hitting the target (since the equation m.v. $\alpha = a$ only signifies that the shell on the average overshoots the target as often as it falls short of it).

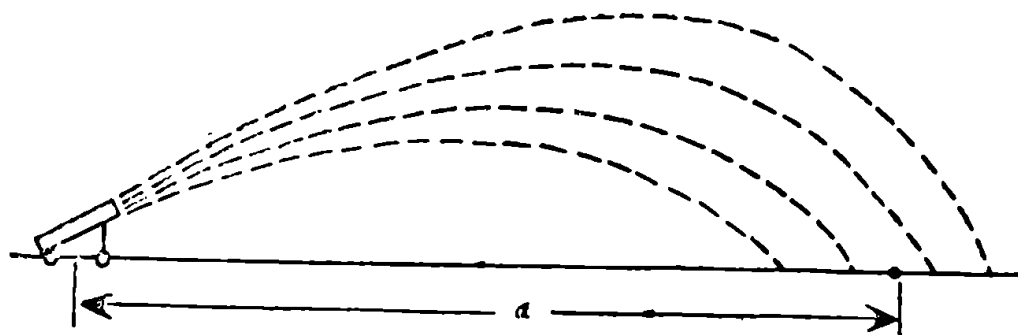


Fig. 3.

How do we determine the accuracy in firing (and compare the performance of two cannons aimed at a target)? The deviation of shells from the target is

given by the number $\alpha - a$; however, the mean value of the random variable $\alpha - a$ is evidently zero:

$$\text{m.v. } (\alpha - a) = \text{m.v. } \alpha - a = a - a = 0,$$

which is obvious since the mean sum of positive and negative values of $\alpha - a$ is zero. It is plain that a nice characteristic of the 'spread' is the mean value of $|\alpha - a|$ (where the vertical lines denote, as usual, the *absolute value* of a number); however, mathematicians do not have much liking for the absolute value of numbers, since it is of little use for further algebraic transformation. Hence, it is usual to characterize the spread of a random variable as *the mean value of the square of its deviation from its mean value*: in fact, the square of both positive and negative numbers is always positive, and no cancellation of the deviations occurs here. The number thus obtained is called the *variance* of the random variable α :

$$\text{Var. } \alpha = \text{m.v. } (\alpha - a)^2 (= \text{m.v. } (\alpha - \text{m.v. } \alpha)^2).$$

The variance of α is the most commonly used measure of the *spread* or *dispersion* (or deviation from the mean value)[†] of α . It is obvious that, in the case of a cannon aimed to strike a target, we consider that cannon to be most appropriate for which the variance of α , the range of flight of the shell, is least (here it is assumed that the cannon is so regulated that the average range of flight of the shell coincides with the distance a from the cannon to the target).

It is easy to comprehend that for the random variable α , characterized by the accompanying probability table

Value	a_1	a_2	\dots	a_k
Probability	p_1	p_2	\dots	p_k

the mean value a is given by

$$a = \text{m.v. } \alpha = p_1 a_1 + p_2 a_2 + \dots + p_k a_k$$

(see, above, p. 6), and the variance is defined by

$$\text{Var. } \alpha = \text{m.v. } (\alpha - a)^2 = p_1(a_1 - a)^2 + p_2(a_2 - a)^2 + \dots + p_k(a_k - a)^2.$$

[†]It is obvious that if, as in the above example, the random variable α has km as its unit of measurement, then its mean value is also measured in km and variance in km². Hence, with variance we frequently consider also a number which is the *square root of the variance* of a random variable. This number is called the *standard deviation* of a random variable:

$$\text{standard deviation of } \alpha = \sqrt{\text{Var. } \alpha};$$

it is measured in the same units as the random variable α and also serves as a measure of the spread of its values.

The last equation can also be set up in a somewhat different form. We note that

$$(\alpha - a)^2 = \alpha^2 - 2a\alpha + a^2.$$

Hence, since the mean value of a sum of (random) variables is the sum of their mean values (see p. 18),

$$\begin{aligned}\text{Var. } \alpha &= \text{m.v. } (\alpha - a)^2 = \text{m.v. } (\alpha^2 - 2a\alpha + a^2) \\ &= \text{m.v. } \alpha^2 + \text{m.v. } (-2a\alpha) + \text{m.v. } a^2.\end{aligned}$$

However, a^2 is not a random variable but a number having a completely definite value†, hence

$$\text{m.v. } a^2 = a^2.$$

On the other hand, the random variable $-2a\alpha$ is obtained from the random variable α by multiplying all its values by $-2a$; hence its mean value also is obtained by multiplying the mean value of α by $-2a$:

$$\text{m.v. } (-2a\alpha) = -2a \times \text{m.v. } \alpha = -2a \times a = -2a^2.$$

Thus, we finally get

$$\begin{aligned}\text{Var. } \alpha &= \text{m.v. } \alpha^2 + \text{m.v. } (-2a\alpha) + \text{m.v. } a^2 = \text{m.v. } \alpha^2 - 2a^2 + a^2 \\ &= \text{m.v. } \alpha^2 - a^2 = \text{m.v. } (\alpha^2) - (\text{m.v. } \alpha)^2,\end{aligned}$$

i.e., *the variance of a random variable is equal to the mean value of its squares minus the square of its mean value.* But, the variance of a random variable is always non-negative (for this is the mean value of the variable $(\alpha - a)^2$, all of whose values are non-negative). It follows from this that the mean value of the square of a random variable is never less than the square of its mean value (see p. 20).

Problem 14. *The accompanying tables of probabilities (frequencies) of the number of defective items (per thousand) is assigned to two identical lathes:*

<i>First lathe : Number of defective items (per thousand)</i>	0	1	2	3	4
<i>Probabilities</i>	0.1	0.2	0.4	0.2	0.1
<i>Second lathe : Number of defective items (per thousand)</i>	0	1	2	3	4
<i>Probabilities</i>	0.15	0.2	0.25	0.3	0.1

†By a^2 we can of course understand a 'random variable' with the accompanying probability table

<i>Values</i>	a^2
<i>Probabilities</i>	1

which implies that

$$\text{m.v. } a^2 = 1 \times a^2 = a^2,$$

i.e., *the mean value of a constant is equal to that constant.*

Compare the mean numbers of defective items produced by the first and the second lathe and the variances of these variables.

It is easy to see that the mean number of defective items produced by the first lathe (α) and those by the second lathe (β) are identical:

$$\text{m.v. } \alpha = 0.1 \times 0 + 0.2 \times 1 + 0.4 \times 2 + 0.2 \times 3 + 0.1 \times 4 = 2,$$

and

$$\text{m.v. } \beta = 0.15 \times 0 + 0.2 \times 1 + 0.25 \times 2 + 0.3 \times 3 + 0.1 \times 4 = 2.$$

From this aspect, both lathes can be regarded to be equivalent. But the variance of the variable α is less than that of β :

$$\begin{aligned} \text{Var. } \alpha &= 0.1 \times (0 - 2)^2 + 0.2 \times (1 - 2)^2 + 0.4 \times (2 - 2)^2 \\ &\quad + 0.2 \times (3 - 2)^2 + 0.1 \times (4 - 2)^2 = 1.2, \end{aligned}$$

and

$$\begin{aligned} \text{Var. } \beta &= 0.15 \times (0 - 2)^2 + 0.2 \times (1 - 2)^2 + 0.25 \times (2 - 2)^2 \\ &\quad + 0.3 \times (3 - 2)^2 + 0.1 \times (4 - 2)^2 = 1.5. \end{aligned}$$

This signifies that production by the first lathe is more 'stable', because here the number of defective items produced in different lots of a thousand items is more densely clustered around the mean value 2 than in the case of second lathe.

We now note that the *variance of the sum of two independent random variables is always equal to the sum of their variances*. In fact, suppose that α and β are two independent random variables, i.e., such that the probability of an individual outcome of one does not at all depend on a particular value taken at that experiment by the other. In this case, as we know (see pp. 15-19), if

$$\text{m.v. } \alpha = a \text{ and m.v. } \beta = b, \text{ then m.v. } (\alpha + \beta) = a + b \text{ and m.v. } (\alpha\beta) = ab.$$

Together with α and β , we consider two more random variables α^2 and β^2 , whose values are the squares of the values of α and β ; for them also we have

$$\text{m.v. } (\alpha^2 + \beta^2) = \text{m.v. } \alpha^2 + \text{m.v. } \beta^2.$$

Further,

$$\text{Var. } \alpha = \text{m.v. } \alpha^2 - a^2; \quad \text{Var. } \beta = \text{m.v. } \beta^2 - b^2,$$

and

$$\begin{aligned} \text{Var. } (\alpha + \beta) &= \text{m.v. } (\alpha + \beta)^2 - [\text{m.v. } (\alpha + \beta)]^2 = \text{m.v. } (\alpha + \beta)^2 - (a + b)^2 \\ &= \text{m.v. } (\alpha^2 + 2\alpha\beta + \beta^2) - (a^2 + 2ab + b^2). \end{aligned}$$

However, since the mean value of a sum of random variables is the sum of their mean values, we have

$$\text{m.v. } (\alpha^2 + 2\alpha\beta + \beta^2) = \text{m.v. } \alpha^2 + \text{m.v. } (2\alpha\beta) + \text{m.v. } \beta^2.$$

But since the random variable $2\alpha\beta$ is twice the random variable $\alpha\beta$, it follows that

$$\text{m.v. } (2\alpha\beta) = 2 \text{ m.v. } (\alpha\beta) = 2ab.$$

Thus, we finally obtain

$$\begin{aligned} \text{Var. } (\alpha + \beta) &= (\text{m.v. } \alpha^2 + 2ab + \text{m.v. } \beta^2) - (a^2 + 2ab + b^2) \\ &= (\text{m.v. } \alpha^2 + \text{m.v. } \beta^2) - (a^2 + b^2) \\ &= (\text{m.v. } \alpha^2 - a^2) + (\text{m.v. } \beta^2 - b^2) = \text{Var. } \alpha + \text{Var. } \beta. \end{aligned}$$

It is obvious that also for an *arbitrary* number of *pairwise independent random variables*, the variance of their sum is equal to the sum of their variances. However, for *non-independent* random variables, this is no longer so. Suppose, for example, that α_1 and α_2 are *one and the same* random variable α with the mean value a , then $\alpha_1 + \alpha_2 = 2\alpha$. In this case, obviously,

$$\text{m.v. } (2\alpha) = 2 \text{ m.v. } \alpha \text{ (i.e., m.v. } (\alpha_1 + \alpha_2) = \text{m.v. } \alpha_1 + \text{m.v. } \alpha_2).$$

However,

$$\text{Var. } (2\alpha) = 4 \text{ Var. } \alpha \text{ (i.e., Var. } (\alpha_1 + \alpha_2) = 2 \text{ Var. } \alpha_1 + 2 \text{ Var. } \alpha_2),$$

since

$$\begin{aligned} \text{Var. } (2\alpha) &= \text{m.v. } [2\alpha - \text{m.v. } (2\alpha)]^2 = \text{m.v. } (2\alpha - 2a)^2 = \text{m.v. } [4(\alpha - a)^2] \\ &= 4 \text{ m.v. } (\alpha - a)^2 = 4 \text{ Var. } \alpha. \end{aligned}$$

Problem 15. A firm manufactures some items, each individual item having the definite probability p of yielding defective pieces (say, $p = 0.002 = 0.2$ per cent). Assuming that all items independently of each other, in some lot of a thousand, are found defective with probability p , find the mean value of the number of defective items per 1000 items produced and the variance of this variable.

We denote by α_i (where $i = 1, 2, 3, \dots$, or 1000) a random variable which assumes the values 1 or 0 according as the i th piece is or is not found defective; in such a case all the 1000 variables have one and the same probability table:

Values	1	0
Probability	p	$1 - p$

Hence,

$$\text{m.v. } \alpha_i = p \times 1 + (1 - p) \times 0 = p (= 0.002),$$

and

$$\begin{aligned} \text{Var. } \alpha_i &= \text{m.v. } \alpha_i^2 - (\text{m.v. } \alpha_i)^2 = [p \times 1 + (1 - p) \times 0] - p^2 = p - p^2 \\ &= p(1 - p) (= 0.002 \times 0.998 = 0.001996). \end{aligned}$$

However, the variable α , we are interested in, is here the sum of all variables α_i , i.e.,

$$\alpha = \alpha_1 + \alpha_2 + \alpha_3 + \cdots + \alpha_{1000}.$$

Moreover, all the variables α_i are mutually independent by hypothesis. Hence,

$$\text{m.v. } \alpha = \text{m.v. } \alpha_1 + \text{m.v. } \alpha_2 + \cdots + \text{m.v. } \alpha_{1000} = 1000 p (= 2),$$

and

$$\text{Var. } \alpha = \text{Var. } \alpha_1 + \text{Var. } \alpha_2 + \cdots + \text{Var. } \alpha_{1000} = 1000 p(1 - p) (= 1.996).$$

The solution of the problem in hand makes use of the fact that *the mean value and the variance of a sum of n mutually independent random variables $\alpha_1, \alpha_2, \dots, \alpha_n$ with the same mean value a and the same variance d are, respectively, equal to n -times the mean value and the variance of a single variable α_1 :*

$$\text{m.v. } (\alpha_1 + \alpha_2 + \cdots + \alpha_n) = n \text{ m.v. } \alpha_1 = na,$$

and

$$\text{Var. } (\alpha_1 + \alpha_2 + \cdots + \alpha_n) = n \text{ Var. } \alpha_1 = nd.$$

In particular, if α is the number of occurrences of a certain event A in a sequence of n mutually independent trials, the probability of the occurrence of A in each trial being p , then

$$\text{m.v. } \alpha = np \text{ and } \text{Var. } \alpha = np(1 - p).$$

From what has been stated above, there emerges a consequence that is quite often useful. We consider the *arithmetic mean*

$$\alpha_m = \frac{\alpha_1 + \alpha_2 + \cdots + \alpha_n}{n}$$

of n mutually independent random variables with the same mean value a and the same variance d . Since all values of the variable α_m are $1/n$ the corresponding values of the variable $\alpha_1 + \alpha_2 + \cdots + \alpha_n$, the mean value of α_m is also $1/n$

times the mean value of the sum $\alpha_1 + \alpha_2 + \dots + \alpha_n$, i.e.,

$$\text{m.v. } \alpha_m = \frac{1}{n} (na) = a.$$

The variance of the random variable α_m is, however, $1/n^2$ times the variance of the variable $\alpha_1 + \alpha_2 + \dots + \alpha_n$ (cf. the assertions above on p. 30 about the variances of α and 2α); hence,

$$\text{Var. } \alpha_m = \frac{nd}{n^2} = \frac{d}{n}.$$

Thus, *the mean value of the arithmetic mean of n mutually independent random variables with the same mean value and variance is equal to the mean value of each of these variables; the variance of the arithmetic mean is, however, less by a factor of $1/n$ the variance of each of the random variables under consideration.*

The inference deduced may now be illustrated by an example. Suppose that we are called upon to determine, to the greatest possible accuracy, the value of some physical quantity a (for concreteness, we may conceive that the question is, say, the determination of some distance on a plane). The result α of a single measurement of the quantity a may be regarded as a *random variable*, since there always exists a definite probability of error due to the inaccuracy of the measuring instrument and carelessness in measurement; in this case, the absence of a *systematic error* in measurement means that

$$\text{m.v. } \alpha = a$$

(cf. p. 26). We now carry out, say 20 independent measurements and form the arithmetic mean α_m of the $\alpha_1, \alpha_2, \dots, \alpha_{20}$ results of these measurements. In this case

$$\text{m.v. } \alpha_m = \text{m.v. } \alpha = a,$$

i.e., the values of the variable α_m , the same as those of α , cluster around the true value a of the measured quantity. However, since

$$\text{Var. } \alpha_m = \frac{1}{20} \text{Var. } \alpha,$$

the spread of the value of α_m is considerably less than that of the values of α . Hence, equating a to the value of α_m , we are fully justified in expecting that a larger error is now considerably *less probable* than in the case in which the result α of a single measurement is actually a . Thus, for instance, if we measure on a plane a distance of the order of 100 m, an error of 1–2 m is often

completely possible; however, the arithmetic mean of 20 independent measurements deviates here almost always from the true value by considerably less than 1 m.

The last remark leads us straight to a remarkable inequality whose derivation is the principal aim of this section. Since $\text{Var. } \alpha_m < \text{Var. } \alpha$, we assume that the probability of a considerable deviation of the variable α_m from its mean value a is less than the probability of a large deviation of α from the number $a = \text{m.v. } \alpha$. This conclusion can be justified rigorously on the basis of the following fundamental result : *if α is a random variable with mean value a and variance d , then we always have*

$$P(|\alpha - a| > \epsilon) < \frac{d}{\epsilon^2}. \quad (*)$$

Here ϵ is an *arbitrary* positive number and the expression $P(|\alpha - a| > \epsilon)$ denotes the probability that the deviation of the value of α from its mean value a is greater than ϵ . The inequality (*) is called *Chebyshev's inequality*; it shows that the less the variance d of the random variable α , the less is the probability of a large deviation of α from the number $a = \text{m.v. } \alpha$.

Chebyshev's inequality (*) is a particular case of another inequality (this also is usually called Chebyshev's inequality), involving an arbitrary random variable β which takes only *non-negative* values. To be precise, *if β takes only non-negative values and the mean value of β is b then, no matter what the positive number c ,*

$$P(\beta > c) < \frac{b}{c}, \quad (**)$$

where $P(\beta > c)$ is the probability that β takes a value greater than c . The inequality (*) obviously follows from (**). In fact, it is only necessary to choose as β the non-negative random variable $(\alpha - a)^2$ whose mean value, by definition, is the variance d of the variable α , and to note that the condition $|\alpha - a| > \epsilon$ is equivalent to the condition $(\alpha - a)^2 > \epsilon^2$; then (**) transforms into (*). Hence, it will suffice for us to prove (**).

Let us assume that the probability table for β has the form

<i>Values</i>	b_1	b_2	b_3	\dots	b_n
<i>Probabilities</i>	p_1	p_2	p_3	\dots	p_n

In this case,

$$b = \text{m.v. } \beta = p_1 b_1 + p_2 b_2 + p_3 b_3 + \dots + p_n b_n.$$

Assume that the possible values of the variable β listed in the above table are

numbered in an ascending order, so that $b_1 < b_2 < b_3 < \dots < b_n$. Of these, suppose that b_k is the first value that exceeds c (i.e., the values b_1, b_2, \dots, b_{k-1} are all less than or equal to c and b_k, b_{k+1}, \dots, b_n are all greater than c); since all values of β are non-negative, the sum on the right-hand side of the preceding equation for b cannot be enlarged if the summands $p_1 b_1 + p_2 b_2 + \dots + p_{k-1} b_{k-1}$ in it are discarded. Consequently,

$$b \geq p_k b_k + p_{k+1} b_{k+1} + \dots + p_n b_n.$$

For all values b_k, b_{k+1}, \dots, b_n on the right-hand side of the last inequality, we now substitute the number c , which is less than these values. Our sum will then get further reduced, and hence

$$b > p_k c + p_{k+1} c + \dots + p_n c = (p_k + p_{k+1} + \dots + p_n) c.$$

Thus, we arrive at the inequality

$$p_k + p_{k+1} + \dots + p_n < \frac{b}{c},$$

which precisely coincides with the inequality (**), since the sum $p_k + p_{k+1} + \dots + p_n$ of the probabilities of those values of β which exceed c is also exactly equal to $P(\beta > c)$.

We now recall the random variable α_m , the arithmetic mean of n independent random variables $\alpha_1, \alpha_2, \dots, \alpha_n$ with exactly the same mean value a and variance d :

$$\alpha_m = \frac{\alpha_1 + \alpha_2 + \dots + \alpha_n}{n}.$$

It is seen from above that

$$\text{m.v. } \alpha_m = a \quad \text{and} \quad \text{Var. } \alpha_m = \frac{d}{n}.$$

Now applying Chebyshev's inequality (*) to the variable α_m , we obtain

$$P(|\alpha_m - a| \geq \varepsilon) < \frac{d}{n\varepsilon^2}. \quad (***)$$

Thus, for instance, suppose that we are given 20 independent measurements spread around a 100 m (such that the average value a of the result of each of these measurements also equals 100 m). Assume that the variance of each measurement is close to 2m^2 . In other words, it is presumed that the squared error of each measurement is on the average equal to 2, i.e., the absolute value of the

error of each measurement is usually of order 1–2 m. In this case, the formula (***) when $\epsilon = 1$ m, yields

$$P(|\alpha_m - 100| > 1) < \frac{2}{20 \times 1^2} = 0.1.$$

Thus, the probability that the arithmetic mean of these 20 measurements deviates from the true value of the distance by more than 1 m is here necessarily less than 0.1.†

We further note especially that if α is the number of occurrences of an event A during n independent trials, the probability of whose occurrence in a single trial is p , then

$$P(|\alpha - np| > n\epsilon) < \frac{np(1-p)}{(n\epsilon)^2},$$

or, equivalently,

$$P\left(\left|\frac{\alpha}{n} - p\right| > \epsilon\right) < \frac{p(1-p)}{n\epsilon^2} \quad (****)$$

[since it is shown on p. 31 that m.v. $\alpha = np$ and $\text{Var. } \alpha = np(1-p)$] for every $\epsilon > 0$. This implies that for every number $\epsilon > 0$ (no matter how small!), we can choose a number n of independent trials so large that the probability

$$P\left(\left|\frac{\alpha}{n} - p\right| > \epsilon\right)$$

of the deviation of the frequency α/n of the realization of the event A in a series of n successive trials from the probability p of the occurrence of A in a single trial by more than ϵ becomes *arbitrarily small*. In fact, for any p and ϵ , the ratio $p(1-p)/n\epsilon^2$ appearing on the right-hand side of inequality (****) tends to 0 as $n \rightarrow \infty$, and this implies that, for sufficiently large n , it is arbitrarily small. But, in real life situations, we usually ignore events having sufficiently small probabilities, regarding them as ‘practically impossible.’ Let us note, however, that the importance of not making a wrong inference influences considerably the decision as to how small the probabilities are that are considered to be small enough to imply that the corresponding events may be taken as impossible ones. Hence, the

†It should also be kept in view that Chebyshev’s inequality (*), as well as the related inequality (***), are highly approximate: the real value of probability at the left-hand sides of these inequalities is most often much less than the corresponding right-hand side. Thus, for example, applying more complex methods, it can be shown that, in the example considered by us, the value $P(|\alpha_m - 100| > 1)$ is actually less than 0.002.

last conclusion means that for every positive ϵ , we can find N so large that the inequality $n > N$ practically guarantees that the deviation of frequency α/n from the probability p is less than ϵ . This consequence, which substantiates the identification of the probabilities of random events with their frequencies, as set forth at the start of this chapter, is called the *law of large numbers* (since it is related to the selection of a large number N of trials).

A similar deduction can be made also from inequality (***), which is more general than inequality (****). Namely, it follows from (***) that for every arbitrarily small positive number ϵ we can always choose a sufficiently large number n of random variables $\alpha_1, \alpha_2, \dots, \alpha_n$ (in other words, a sufficiently large number of observations or trials), such that it guarantees us that the probability $P(|\alpha_m - a| > \epsilon)$ will be sufficiently small. In fact, for every ϵ (and every fixed value of d) the right-hand side $d/n\epsilon^2$ of inequality (***) also tends to zero as n increases indefinitely. Thus, for every $\epsilon > 0$, we can, by means of the choice of a sufficiently large number n , guarantee the practical reliability of the inequality $|\alpha_m - a| < \epsilon$. The general statement that, for a sufficiently large number of similar independent trials (i.e., independent trials leading to numerical results that have the same mean value and variance), the arithmetic mean of their results $\alpha_1, \alpha_2, \dots, \alpha_n$ can be made as close as desired to the mean value a of the variables $\alpha_1, \alpha_2, \dots, \alpha_n$, is also called the *law of large numbers*.

In fact, we may even dispense with the requirement that the mutually independent random variables $\alpha_1, \alpha_2, \alpha_3, \dots$ involved in the determination of the quantity α_m should have the same mean value and variance. Indeed, if the mean values of these random variables are a_1, a_2, a_3, \dots and their variances d_1, d_2, d_3, \dots are bounded (i.e., there is a number D such that $d_i < D$ for every i), then from Chebyshev's inequality (*) it follows that

$$P(|\alpha_m - a_m| > \epsilon) < \frac{D}{n\epsilon^2}, \quad \text{where } a_m = \frac{a_1 + a_2 + \dots + a_n}{n}.$$

This implies in turn that for every number $\epsilon > 0$ we can, by choosing a sufficiently large number n , practically guarantee that the inequality $|\alpha_m - a_m| < \epsilon$ is satisfied. This assertion is also another form of the *law of large numbers*.

1.5. Algebra of events and general definition of probability

In the earlier section, a key role is played by two operations, which associated the two events A and B and a certain third event; we have designated these operations as the *sum* and *product* of A and B , written $A + B$ and AB (see pp. 8 and 11). Some justification for these names is provided by the fact that the rules of 'addition' and 'multiplication' of events strongly remind us of the rules of addition and multiplication of numbers. Thus, from the very definition of the sum and product of two events it follows that $A + B = B + A$ and $AB = BA$; at one place we also made use of the equality $(A + B)C = AC + BC$ (see p. 14).

The present section aims to analyse more closely the points of similarity and dissimilarity between the 'algebra of events' and the 'algebra of numbers'.

In arithmetic and algebra, we consider various sorts of numbers e.g. integers, rational, real (both rational and irrational) and complex numbers. In every case, with each pair of numbers a and b , there can be associated two other numbers, their sum $a + b$ and product ab . In addition, the rules involving addition, closely resemble the rules involving multiplication. Thus, for instance :

$$\begin{array}{lll} a + b = b + a & \text{and} & ab = ba \\ (a + b) + c = a + (b + c) & \text{and} & (ab)c = a(bc). \end{array}$$

This analogy between the operations of addition and multiplication is also manifested in the existence of two *idempotent* numbers 0 and 1 such that the addition of the former and the multiplication by the latter do not alter any number:

$$a + 0 = a \text{ and } a \times 1 = a.$$

This analogy, unfortunately, does not go very far. The reason is the asymmetric distributive law

$$(a + b)c = ac + bc,$$

where addition and multiplication appear in entirely different roles. Indeed, if, in the last equation, the addition sign is replaced everywhere by the multiplication sign and vice versa, we arrive at the absurd 'equality'

$$a \times b + c = (a + c) \times (b + c).$$

Hence, many properties of addition and multiplication are very different from each other. Thus, for instance, the number 0 plays a most critical role in relation to multiplication, as it is seen from the important equality

$$a \times 0 = 0$$

(which implies, in particular, that the division of a number a , different from zero, by 0 is impossible); in contrast to this the following analogue of the above equality involving addition does not obviously hold:

$$a + 1 = 1.$$

However, there also exist objects, other than numbers, for which we can define the operation of addition and multiplication, sharing many properties of the addition and multiplication of numbers. In some of these cases, we obtain algebraic systems, where a greater closeness than the one in the case of numbers prevails between the operations of addition and multiplication defined in these systems. Let us, for example, consider a collection of all possible sets (or 'figures') of a plane. The *sum* $A + B$ of two sets A and B is naturally defined as their *union* (Fig. 4a). Then, we obviously have

$$A + B = B + A,$$

and

$$(A + B) + C = A + (B + C)$$

(in the last equation the union of three sets A , B and C appears on the left- and right-hand sides, which can also be written simply as $A + B + C$, without parentheses). The role of zero

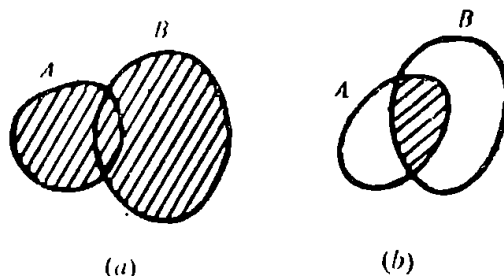


Fig. 4.

is played here by the so-called 'empty' set O , which contains no point; for such a set we have

$$A + O = A.$$

We now define the *product* AB of two sets A and B as their common part or *intersection* (Fig. 4b). It is obvious, in this case, that

$$AB = BA,$$

and

$$(AB)C = A(BC)$$

(in the last equation the common part of three sets A , B and C occurs on the left- and right-hand sides, and it is natural to denote it simply ABC). The role of unity is played here by the entire plane I . Indeed, for every set A we have

$$AI = A.$$

It is easy to show that for an algebra of sets so defined, the distributive law

$$(A + B) \times C = A \times C + B \times C$$

holds. To prove this law it suffices to consider Fig. 5a, where sets $A + B$ and C are distinguished by two different types of shading, so that their product (intersection) $(A + B)C$ is shaded by double strokes; I denotes the product $A \times C$ and II the product $B \times C$. However, we have here the 'second distributive law'

$$A \times B + C = (A + C) \times (B + C),$$

which is obtained from the first one by interchanging the roles of addition and multiplication. For the proof of this, it suffices to consider Fig. 5b, where the sets $A + C$ and $B + C$ are shaded in two different ways, so that their product $(A + C) \times (B + C)$ is shaded with double strokes; the portion I denotes the set $A \times B$ and II the set C .

The duality between these two distributive laws completely defines the analogy between the rules involving the addition and multiplication of sets. Thus, for instance, it is obvious here that

$$A \times O = O \quad \text{and} \quad A + I = I.$$

We can compare also the two equations

$$A \times A = A \quad \text{and} \quad A + A = A,$$

none of which holds in the algebra of numbers.

In arithmetic and algebra, an important role is played by the comparison of numbers according to their magnitude. If we regard \leq to be the basic sign of comparison (the relation

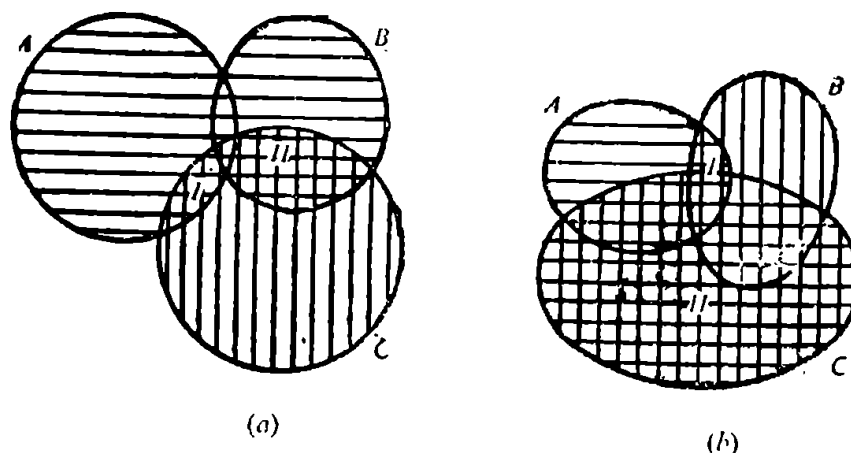


Fig. 5.

$a \leq b$ means that the number a is *not greater* than the number b), then the basic rules of comparison of numbers take the following form:

- | | | |
|--|------------|--|
| | $a \leq a$ | (every number a is not greater than itself); |
| if $a \leq b$ and $b \leq a$, then $a = b$ | | (if a is not greater than b and b is not greater than a , then a and b are equal); |
| if $a \leq b$ and $b \leq c$, then $a \leq c$ | | (if the number a is not greater than b and b is not greater than c , then a is not greater than c). |

We can also introduce here the *comparison of sets*, conventionally written as $A \subset B$ (the sign \subset replaces the 'composite' sign \leq), if A is a part of the set B (A can also coincide with the whole of B). Here, it is also obvious that†

$$\begin{aligned} A &\subset A; \\ \text{if } A &\subset B \text{ and } B \subset A, \text{ then } A = B; \\ \text{if } A &\subset B \text{ and } B \subset C, \text{ then } A \subset C. \end{aligned}$$

Among other set-comparison rules, the following are noteworthy:

$$A \subset A + B \quad \text{and} \quad AB \subset A,$$

and also

$$A \subset I \quad \text{and} \quad O \subset A.$$

(The last relation says that an empty set O contains no point other than a point of the set A . This is true for *every* A , since O has no point in it.)

A salient difference between the algebra of sets and the algebra of numbers consists in the concern of the former with an additional operation, which puts in correspondence with every set A , a new set \bar{A} (the *complement* of A). This operation is defined as follows: \bar{A} consists

† We note one substantial difference between number and set comparisons. For every pair of (real) numbers a and b , one of the two relations $a \leq b$ and $b \leq a$ is necessarily valid (even both may be valid if a and b are equal). In contrast to this, for a pair of sets A and B , *none* of the relations $A \subset B$ and $B \subset A$ is satisfied on many occasions. (A similar situation holds also for *complex* numbers, if it is agreed upon, which happens in some investigations, to write $a \leq b$ when the complex numbers a and b have the same argument and the modulus of a does not exceed that of b .)

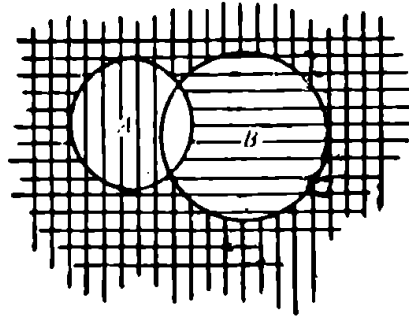


Fig. 6.

of all the points of a plane, which *do not belong* to A . The main rules of this new operation are

$$\begin{aligned} A + \bar{A} &= I \quad \text{and} \quad A\bar{A} = O; \\ \bar{O} &= I \quad \text{and} \quad \bar{I} = O; \\ \bar{\bar{A}} &= A; \end{aligned}$$

$$\text{if } A \subset B, \text{ then } \bar{B} \subset \bar{A};$$

and, finally,

$$\overline{A + B} = \bar{A} \times \bar{B} \quad \text{and} \quad \overline{A \times B} = \bar{A} + \bar{B}$$

(see Fig. 6, in which the sets \bar{A} and \bar{B} are shown with different types of shading, the set $\overline{A + B}$ is doubly shaded and the set $\overline{A \times B}$ has at least a single shading).

There are also many other collections of some objects for which it is natural to define the concepts of sum, product, and even the 'ordering' $A \subset B$ as well as the 'complement' \bar{A} , which satisfy all the algebraic properties enumerated above. One example of this class is a collection of random events considered in Secs. 1-3; as is easy to see, all the properties of set algebra carry over to the algebra of events. Another example can be obtained if, instead of a set of points on a plane, we consider a set of elements of some other nature, say, a set of integers. If, in addition, by the sum and product of the sets A and B we understand, as before, their union and intersection (for example, if A_2 and A_3 are sets of numbers divisible by 2 and by 3, respectively, then the set $A_2 + A_3$ contains all even numbers and those odd numbers which are divisible by 3, while the set $A_2 A_3$ consists of all integers which are multiples of 6) and regard that $A \subset B$ if A forms a part of B (say $A_4 \subset A_2$, where A_4 is a set of numbers divisible by 4) and that \bar{A} is a set of all integers *not belonging to* A (if A is a set of all prime numbers, then \bar{A} contains all composite numbers and the number 1), and I and O are taken, respectively, as a set of *all* integers and a set that has not a *single* number in it, then all the relations enunciated above remain valid.

As one more example, we can consider a set of all divisors of a certain number N , which is not divisible by any square greater than 1 (to be specific, for $N = 30$, a set of numbers 1, 2, 3, 5, 6, 10, 15, 30); if by $A + B$ and AB we understand, respectively, the *least common multiple* and the *greatest common divisor* of numbers A and B , by $A \subset B$ the relation that ' A is a divisor of B ' and denote by O and I the numbers 1 and N (i.e., 1 and 30) and by \bar{A} the number N/A

(in our case $30/A$) then, as before,

$$\begin{aligned} A + B &= B + A \quad \text{and} \quad AB = BA, \\ (A + B) \times C &= A \times C + B \times C \quad \text{and} \quad A \times B + C = (A + C) \times (B + C) \\ A &\subset A + B \quad \text{and} \quad AB \subset A, \\ \overline{A + B} &= \overline{A} \times \overline{B} \quad \text{and} \quad \overline{A \times B} = \overline{A} + \overline{B}, \end{aligned}$$

and so on.

Finally, the most important example in this direction is furnished by the set of all logical propositions (i.e., all statements such that each has a meaning when it says that a proposition is true or false); this set is the object of study in *mathematical logic*. Here, by the sum $A + B$ and the product AB one ought to understand the statements 'either A or B ' and 'both A and B ', respectively; by $A \subset B$, the fact that the truth of A implies also the truth of B (for short ' A implies B '); by \overline{A} , the negation of A (the proposition ' A is not true'); and by I and O the propositions which are a fortiori true and false, respectively. In this case again, all the relations described above are satisfied, which express the definite laws of logic. Thus, for example,

$$A + \overline{A} = I$$

is the *law of the excluded middle*: in all cases, the proposition A is either true or false; the relation

$$A \times \overline{A} = O$$

is the *law of contradiction*: the proposition A cannot be simultaneously both true and false.

The versatility and importance of algebraic systems that possess all the properties enumerated above motivated mathematicians to study them especially. At present, such systems are called 'Boolean Algebras'†, named after George Boole, the celebrated English mathematician and logician of the nineteenth century, who was first to apply such an algebra in his researches in the field of logic.

The elements of a Boolean algebra are generally not numbers. However, we often succeed, in associating with every element A , the number $|A|$ or $p(A)$ satisfying the following conditions

$$\begin{aligned} 0 &\leq p(A) \leq 1; \quad p(O) = 0, \quad p(I) = 1; \\ \text{if } A &\subset B, \quad \text{then } p(A) \leq p(B); \\ \text{if } A \times B &= O, \quad \text{then } p(A + B) = p(A) + p(B). \end{aligned}$$

This number is called the *absolute value* or *norm* of A and the Boolean algebra itself in this case is called a *normed algebra*. By way of an example, we may mention a family of plane figures, belonging to a square with unit side (the square itself plays the role of the element I of this Boolean algebra), where area of Fig. A is taken as the absolute value or norm of A . Other

† A Boolean algebra can be characterized as a collection of elements, where two operations \overline{A} and $A + B$ are defined (associating with every element A , respectively a pair of elements \overline{A} and B , some element of the same set), having the following properties

$$\begin{aligned} A + B &= B + A, \\ (A + B) + C &= A + (B + C), \\ \overline{\overline{A} + \overline{B}} + \overline{\overline{A} + B} &= A. \end{aligned}$$

All the remaining properties of Boolean algebra can be derived from these three basic properties, if we define the 'product' AB as $\overline{\overline{A} + \overline{B}}$, the relation $A \subset B$ as the equality $A + B = B$, the elements I and O as the right-hand sides of equations $A + \overline{A} = I$ and $A\overline{A} = O$ (A being arbitrary).

examples are a set of all divisors of integer N , where N does not divide the square of any integer (N may be, for example, a number 30); here by the norm of A is understood $\log_N A$ (in other words, $\log_{30} A$). A collection of all propositions of mathematical logic can also be treated as a normed Boolean algebra if it is agreed to regard the absolute value (norm) of a proposition to be 1 if it is true, or 0 if it is false. An example of normed Boolean algebra is also the algebra of events, studied in Secs. 1—3; the role of absolute value or norm of an event A is played here by the probability $p(A)$ of this event.

The link between probability theory and Boolean algebra can be used as the foundation stone for the general definition of the subject matter of probability. Namely, we can assert that *the theory of probability studies a collection of objects, which form a normed Boolean algebra*; these objects are called *events* and the norm $p(A)$ of an event A is called its *probability*. Thus, for example, in the 'urn problem' (or in every problem reducible to it) we actually consider a Boolean algebra of all possible sets which can be composed of n given elements (points). In addition, the sum and product of two sets (as also in all the examples below) are defined as their union and intersection; the norm is, however, defined by the condition that for every single element (i.e., isolated point) set it is equal to one and the same number $1/n$. However, so very legitimate, from our new view-point, are the problems that arise from invoking the same Boolean algebra, though under more general conditions, that we equate the norms of isolated points with arbitrary positive numbers p_1, p_2, \dots, p_n , which satisfy the unique condition $p_1 + p_2 + \dots + p_n = 1$ (in particular, the problem of an imperfect die having a distorted form or having been made of inhomogeneous material reduces to a Boolean algebra of such type with $n = 6$). We shall encounter later also a case in which the elements of a Boolean algebra are all possible parts of a given segment AB , but the norm is defined as the ratio of the length of the part under consideration to the entire length of the segment AB (see Problem 22, Chap. 2). Quite similarly, it is sometimes useful to consider a collection of all sets belonging to some plane figure or spatial body and define the norm as the ratio of the area or volume of the corresponding set to the area of the entire figure or volume of the entire body (see, for example, 'Experiments with infinitely many possible outcomes' on pp. 27-30 in [40]). The 'problem of an imperfect die' can also be generalized to all these cases, i.e., even when considering a Boolean algebra of all sets belonging to a given segment, or figure of a body, we can introduce a norm in a completely arbitrary manner with the only requirement being that it satisfy the conditions imposed above on the function $p(A)$. We thus arrive at a new wide class of interesting probability-theoretic problems.

If the italicized statement in preceding paragraph is taken as the definition of probability, then it implies that, in every problem related to this theory, the basic Boolean algebra must necessarily be determined beforehand (i.e., it must be indicated in one way or another in the conditions of the problem itself). The main problem of the theory of probability should then be regarded as the determination of the probabilities of compound events formed of the given basic or elementary events A, B, C, D, \dots by means of the operations of Boolean algebra (for example, of the event $AB + BC + CA$ or $(A + B \times C)(A + D)$) when the probabilities of these elementary events are considered to be known (just as the main problem in geometry consists of the calculation of some length or angles with respect to other original lengths or angles, assumed to be known; for example, the length of the hypotenuse of a right triangle with respect to the lengths of two legs of this triangle). In such an approach to probability theory (indicated first by the Russian mathematician S. N. Bernstein in 1917) the crucial problem of the methods of evaluating the basic probabilities $p(A)$, $p(B)$, and so on, obviously remains open. However, the developing theory will have a practical value, only if these probabilities can be determined in such a way that they coincide with the empirical frequencies of the corresponding events in a long series of experiments. One possible way to determine the 'basic probabilities' satisfying this condition is given by the 'classical definition of probability' adduced in Sec. 1, which rests on the concept of the 'complete system of equally probable

outcomes of an experiment' (this 'classical definition' was first introduced by P. S. Laplace). In other cases when such a complete system does not exist, we take recourse to different routes for determining the values of $p(A)$; for example, via finding the approximate value of $p(A)$ directly by means of the repeated performance of an experiment to which the occurrence of the event A is related. The heart of the matter, however, is that the methods of determining the original probabilities are not at all reflected in all the succeeding operations over them, which form the main content of the theory.

We also note the situation that, in all the examples set forth above, we define Boolean algebra as a collection of sets composed of the points of a 'super set'. This circumstance is not accidental; it is possible to show that such formulation of this algebra is possible in *all* probabilistic problems. Proceeding from this, one can reckon even from the start that the basic object of study of the probability theory is not the normed Boolean algebra of all the possible events but a 'set of all possible elementary events' whose various parts (subsets) are later identified as the 'events'. In order to bring these arguments to their logical conclusion, it is simply necessary to assign a well-defined norm $p(A)$ to a subset A of our 'set of elementary events' and prescribe the main requirements (axioms) which must be satisfied by the subsets under consideration and their norms, so that indeed we have a normed Boolean algebra. This approach to the axiomatic construction of probability theory (proposed by A. N. Kolmogorov in 1929-1932) has definite advantages over the method shown above in this section for investigating more complex and subtler questions of the probability theory. Therefore, this approach gained the widest popularity in modern times and is now most extensively used. However, we refrain from a deeper involvement with this topic in order not to be led far away from the main theme of the book.

2

Entropy and information

2.1. Entropy as a measure of the amount of uncertainty

The main property of random events, whose study is the basic object of this book, is a complete lack of confidence in their occurrence, which creates the well-known uncertainty about the outcomes of an experiment related to these events. However, it is fully obvious that the amount of this uncertainty is different in different cases. If our experiment consists of determining the colour of the first raven which we will see, then, of course, we can with almost absolute confidence consider it to be black. In fact, though ornithologists say that, in principle, white ravens are also existing, hardly anyone will entertain a doubt about the outcome of such an experiment. Somewhat less certain is an experiment which consists of ascertaining whether or not the first person, we collide with, will be left-handed; here also one can predict the result of the experiment without any hesitation, but the risk to fail in this prediction is still greater than that in the first case. It is considerably more difficult to predict beforehand whether the first person whose path we will cross in the street of a city will be a male or a female. But even this experiment has relatively smaller uncertainty as compared to, say, the one of indicating in advance who will be the winner in a tournament with twenty participants completely unknown to us, or what will be the number of the lottery ticket which will win first prize in a forthcoming draw. If we predict, say, that the first person we meet in the street will be a male, we still have a hope for the success of our conjecture, but hardly anyone will hazard a forecast in the penultimate or much less in the last case.

For practical purposes, it is important to know how to evaluate the *degree of uncertainty* of highly diverse experiments, in order that we may have an opportunity to compare them from this aspect. To start with, let us consider the experiments that have k *equally likely* outcomes. It is obvious that the degree of uncertainty of each such experiment is determined by the number k : if, for $k = 1$, the outcome of an experiment is not random at all, then for k large, i.e., when a large number of different outcomes is involved, a forecast of the result of the experiment becomes very difficult. It is thus quite clear that the desired numerical measure of uncertainty must depend on k , i.e., it must be a function $f(k)$.

In addition, for $k = 1$, this function must reduce to zero (because in this case there is no uncertainty), and it must increase with increasing k .

For a fuller definition of the function $f(k)$, it is necessary to impose additional restrictions on it. We consider two *independent* experiments α and β (i.e., two experiments such that the outcomes of one has no effect on the probabilities of the outcomes of the other). Suppose that α and β have, respectively, k and l equally probable outcomes; consider also a compound experiment $\alpha\beta$ consisting of the simultaneous occurrence of α and β . It is obvious that the uncertainty of $\alpha\beta$ is larger than that of α , since here the uncertainty of the outcome of β is added to that of α . It is, therefore, natural to assume that the *degree of uncertainty of $\alpha\beta$ equals the sum of the uncertainties characterizing experiments α and β* . But since the experiment $\alpha\beta$ has obviously kl outcomes of equal probability (they are obtained by combining each k of the possible outcomes of α with the l outcomes of β), we arrive at the following condition which must be satisfied by the function $f(k)$:

$$f(kl) = f(k) + f(l).$$

From the last condition stems the suggestion that we take *the number $\log k$ as a measure of uncertainty of an experiment that has k outcomes of equal probability* (because $\log(kl) = \log k + \log l$). Such a definition of the measure of uncertainty also agrees with the conditions that it is equal to 0 for $k = 1$ and that it increases with increasing k .†

We note that the choice of a base for the system of logarithms is immaterial here since by virtue of the well-known formula

$$\log_b k = \log_a a \times \log_a k$$

a transition from one system of logarithms to another reduces to only the multiplication of the function $f(k) = \log k$ by a constant factor (*the factor of transition $\log_a a$*). In other words, such a transition is equivalent merely to a change in the *unit of measurement* of the amount of uncertainty and is, therefore, fundamentally a matter of indifference. In specific applications of a 'measure of the uncertainty' it is customary to use logarithms to the base 2 (in other words, to consider that $f(k) = \log_2 k$). This means that we choose here, as a unit of the uncertainty, the uncertainty of an experiment that has *two* outcomes of equal probability (say, flipping of a coin to determine a 'head' or 'tail', or finding out the answer 'yes' or 'no' to a question apropos of which we can expect with equal justification the answer to be affirmative or negative). Such a unit of measurement of uncertainty is called a *binary unit* (abbreviated to *bit*); in the

†It is easy to show that a logarithmic function is a *unique* function of the argument k , which satisfies the conditions $f(kl) = f(k) + f(l)$, $f(1) = 0$ and $f(k) > f(l)$ for $k > l$ (see Sec. 4 below).

German literature it is known also by a more expressive 'Ja-Nein Einheit' (yes-no unit). Such a 'yes-no unit' is in a certain sense most natural; Chapter 4 will further elaborate upon the considerations that led to its adoption in engineering. We shall also use binary units (bits) throughout in what follows; thus, the expression $\log k$ (where, as a rule, the base of the system of logarithms is omitted) usually denotes $\log_2 k$ in this book. It is, however, worth noting that in the content of this book there would be practically no change if we were to use the more common *decimal logarithms*; this would only imply the choice of a unit for the measurement of uncertainty of an experiment that has 10 outcomes of equal probability (such, for example, is an experiment that consists of drawing a ball from an urn with ten numbered balls or an experiment involving the finding of a digit if each of the ten digits were to have the same probability of being thought of). This last unit for the measurement of uncertainty (called the *decimal unit or dit*) is roughly $3\frac{1}{3}$ times greater than the binary unit (since $\log_2 10 \approx 3.32 \approx 3\frac{1}{3}$).

The probability table for an experiment that has k equally likely outcomes has the form

<i>Outcomes of experiment</i>	A_1	A_2	A_3	\dots	A_k
<i>Probabilities</i>	$\frac{1}{k}$	$\frac{1}{k}$	$\frac{1}{k}$		$\frac{1}{k}$

Since we agree that the total uncertainty of such an experiment is $\log k$, it can be considered that every individual outcome with the probability $1/k$ introduces an uncertainty equal to $1/k \log k = -1/k \log 1/k$. But, then, it is natural to regard that in the case of an experiment with the probability table

<i>Outcomes of experiment</i>	A_1	A_2	A_3
<i>Probabilities</i>	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{6}$

the outcomes A_1 , A_2 and A_3 introduce uncertainties, which are, respectively, equal to $-\frac{1}{2} \log \frac{1}{2}$, $-\frac{1}{3} \log \frac{1}{3}$ and $-\frac{1}{6} \log \frac{1}{6}$. If so, then the total uncertainty of this experiment is given by

$$-\frac{1}{2} \log \frac{1}{2} - \frac{1}{3} \log \frac{1}{3} - \frac{1}{6} \log \frac{1}{6}.$$

Quite similarly, we can assert that in the most general case, *for an experiment α with the probability table*

Outcomes of experiment	A_1	A_2	A_3	\dots	A_k
Probabilities	$p(A_1)$	$p(A_2)$	$p(A_3)$	\dots	$p(A_k)$

the measure of uncertainty is given by

$$-p(A_1) \log p(A_1) - p(A_2) \log p(A_2) \\ - p(A_3) \log p(A_3) - \dots - p(A_k) \log p(A_k)$$

(see also Sec. 4 of this chapter in small print). We call this last number the *entropy* of an experiment α and denote it by $H(\alpha)$, thus following a deep physical analogy which there is no need to go into here.†

We now study the properties of the entropy $H(\alpha)$. We note in the first place that it cannot take negative values: since we always have $0 \leq p(A) \leq 1$, it follows that $\log p(A)$ cannot be positive, and hence $-p(A) \log p(A)$ cannot be negative. We further note that, if p is very small, then the product $p \log p$ is also quite small, even though $-\log p$ is here a large positive number. In fact, for example, let $p = 1/2^n$; then $\log p = -n$ and $-p \log p = n/2^n$. It is clear that the fraction $n/2^n$ for large n (which corresponds to small $p = 1/2^n$) is quite small (because with increasing n the number 2^n grows much faster than n itself; thus, for example, the number 2^{64} consists of 20 digits)!†† Hence, it follows that as $p \rightarrow 0$ the product $-p \log p$ decreases unboundedly, so that

$$\lim_{p \rightarrow 0} (-p \log p) = 0$$

(cf. Figs. 7 and 9 depicting the graph of the function $y = -p \log p$; it is seen from these graphs that when $p = 0$ the value of this function is 0). Hence, if the probability $p(A_i)$ of the outcome A_i is zero (i.e., the outcome A_i is impossible), then the corresponding term $-p(A_i) \log p(A_i)$ in the expression for entropy can be discarded without any qualms (strictly speaking, this term makes no sense, since $\log p(A_i)$ in this case does not exist; just because of this we take recourse to finding the *limit* of the expression $-p \log p$ as $p \rightarrow 0$). Contrarily, when $p(A_i)$ is quite large (i.e., close to 1), the term $-p(A_i) \log p(A_i)$ is also

†If we relate the concept of entropy introduced here to the thermodynamic concept of entropy, it plays an important role in physics; see, for example, Brillouin [5] (cf. also Poletayev [18]).

††Many readers may be aware of a legend related to this, which states that the inventor of chess, when asked to name his reward, requested as many grains of food as would result from putting one grain on the first square on the board, two on the second and then on each succeeding square double the number of grains on the preceding one. This reward, as reckoned initially by its squares (64), was envisioned to be quite modest; however, the corresponding number of grains (equal to $2^{64} - 1$) actually far exceeded the entire stock of food grain on earth.

quite small, since $\log p$ tends to zero as $p \rightarrow 1$. If the probability $p(A_i)$ is precisely 1 (i.e., the occurrence of the outcome A_i of our experiment is the certain event), then $\log p(A_i) = 0$ and hence also $-p(A_i) \log p(A_i) = 0$ (see again Figs. 7 and 9).

Since $-p \log p$ is 0 if and only if $p = 0$ or $p = 1$, it is clear that the *entropy* $H(\alpha)$ of an experiment α is 0 if and only if one of the probabilities $p(A_1), p(A_2), \dots, p(A_k)$ is 1 and all the others are 0 (recall that $p(A_1) + p(A_2) + \dots + p(A_k) = 1$; see p. 9 above). This situation agrees well with the purport of the quantity $H(\alpha)$ as a measure of the uncertainty; in reality, it is only in this case that there is no uncertainty about an experiment.

Furthermore, it is natural to consider that, among all experiments having k outcomes, an experiment α_0 with the following probability table is *most uncertain*:

Outcomes of experiment	A_1	A_2	A_3	\dots	A_k
Probabilities	$\frac{1}{k}$	$\frac{1}{k}$	$\frac{1}{k}$	\dots	$\frac{1}{k}$

In fact, in this case it is most difficult to predict the outcome of the experiment. This corresponds to the circumstance that the experiment α_0 has the *largest* entropy: if α is an arbitrary experiment with k outcomes A_1, A_2, \dots, A_k , then

$$H(\alpha) = -p(A_1) \log p(A_1) - p(A_2) \log p(A_2) - \dots - p(A_k) \log p(A_k) \\ \leq \log k = \underbrace{-\frac{1}{k} \log \frac{1}{k} - \frac{1}{k} \log \frac{1}{k} - \dots - \frac{1}{k} \log \frac{1}{k}}_{k\text{-times}} = H(\alpha_0),$$

where equality is valid if and only if $p(A_1) = p(A_2) = \dots = p(A_k) = 1/k$. We defer a complete proof of this conclusion for the present (see Appendix I at the end of the book); here, however, we confine ourselves to illustrating the

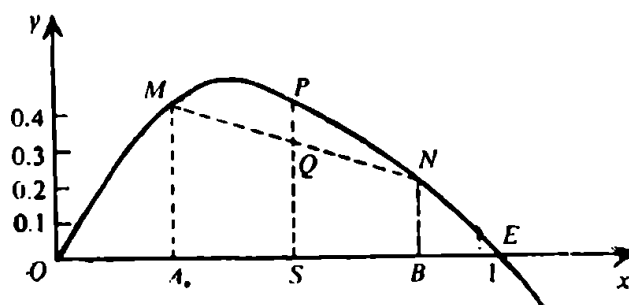


Fig. 7.

related theorem by an example in which $k = 2$. In this case, the theorem reduces to proving the inequality

$$-p(A_1) \log p(A_1) - p(A_2) \log p(A_2) \leq \log 2 = 1,$$

where

$$p(A_2) = 1 - p(A_1).$$

As already remarked, the value of the function $F(x) = -x \log x$ tends to zero as $x \rightarrow 0$; on the other hand, for $x = 1$ also its value is zero, and for $0 \leq x \leq 1$ this function is positive (because in this case $\log x$ is negative); for $x > 1$ the function $-x \log x$ is negative. The graph of the function under consideration is shown in Fig. 7, where $OE = 1$, $OA = p(A_1)$, $OB = p(A_2)$ and the segments AM and BN depict the variables $-p(A_1) \log p(A_1)$ and $-p(A_2) \log p(A_2)$. Since

$$OA + OB = p(A_1) + p(A_2) = 1 = OE,$$

the distance OS from the origin to the center S of the segment AB equals $\frac{1}{2}$; hence in Fig. 7, the segment SP equals $-\frac{1}{2} \log \frac{1}{2} = \frac{1}{2}$. But the half sum of the segments AM and BN equals the middle line SQ of the trapezium $ABNM$, which does not exceed SP ; consequently,

$$\frac{1}{2} (-p(A_1) \log p(A_1) - p(A_2) \log p(A_2)) \leq \frac{1}{2},$$

i.e.,

$$-p(A_1) \log p(A_1) - p(A_2) \log p(A_2) \leq 1,$$

where the equality holds if and only if the segments OA and OB both coincide with OS . Thus, it is shown that the function

$$h(p) = -p \log p - (1 - p) \log (1 - p),$$

which defines the entropy of an experiment with two outcomes (whose probabilities are p and $1 - p$), assumes its largest value (i.e., $\log 2 = 1$) when $p = \frac{1}{2}$.

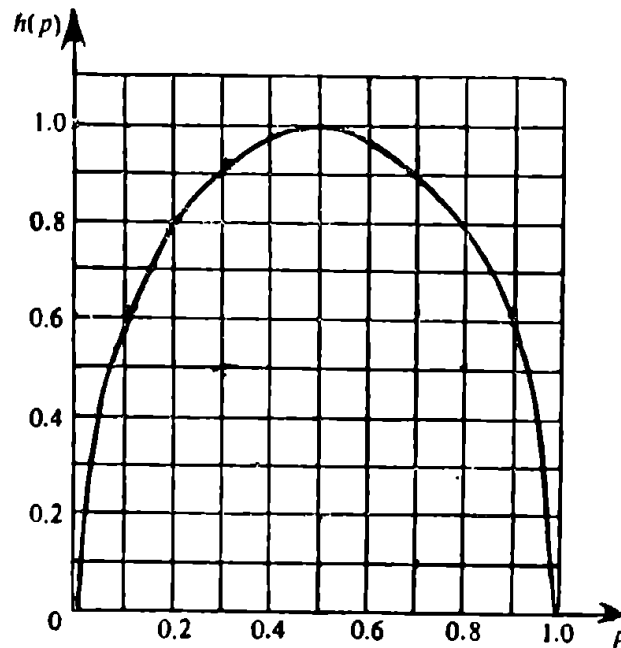


Fig. 8.

The graph of this function is given in Fig. 8, which shows how the entropy $h(p)$ varies for p varying between 0 and 1.[†]

In the case of an experiment with k possible outcomes, the entropy is given by the formula

$$H(p_1, p_2, \dots, p_k) = -p_1 \log p_1 - p_2 \log p_2 - \dots - p_k \log p_k,$$

where p_1, p_2, \dots, p_k are the probabilities of individual outcomes, so that we always have $p_1 + p_2 + \dots + p_k = 1$. This is a generalisation of a case considered above (because, when $k = 2$, the function $H(p_1, p_2, \dots, p_k)$ turns into $H(p_1, 1 - p_1) = h(p_1)$); it can also be shown that the function $H(p_1, p_2, \dots, p_k)$ assumes its largest value, namely, $\log k$, when $p_1 = p_2 = \dots = p_k = 1/k$; for the proof, see Appendix I. In order to bring out the nature of the relationship between the function $H(p_1, p_2, \dots, p_k)$ and the individual probabilities p_1, p_2, \dots, p_k , we consider again a graph of the function $-p \log p$, $0 < p < 1$ (see Figure 9, where a part of Figure 7 is reproduced to a somewhat larger scale).

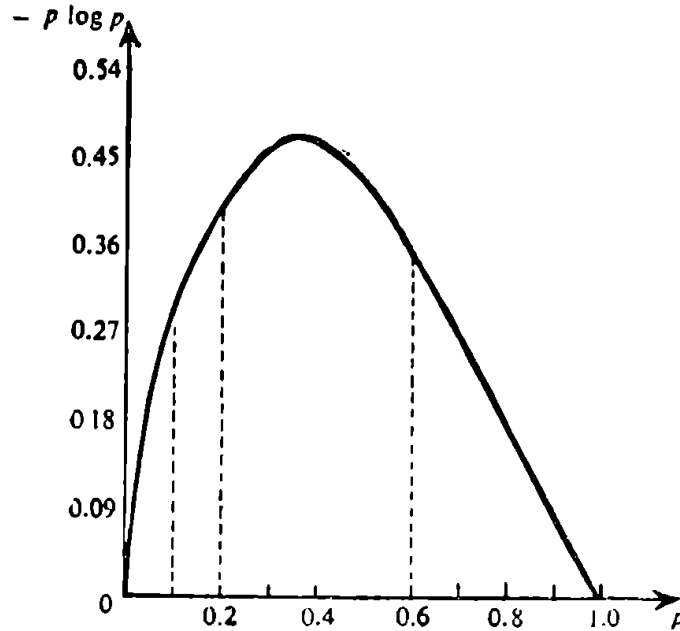


Fig. 9.

From this graph it is seen that, when $p < 0.1$, the quantity $-p \log p$ grows extraordinarily fast; hence in this range a comparatively small decrease in the probability p_i results in a highly significant decrease of the corresponding term $-p_i \log p_i$ in the expression of the function $H(p_1, p_2, \dots, p_k)$. This leads us to the fact that the summands $-p_i \log p_i$, which correspond to *very small*

[†]Tables of values of the functions $-p \log p$ and $h(p) = -p \log p - (1 - p) \log (1 - p)$ (logarithms are binary ones) are given in Appendices III and IV of this book.

values of the probability p_i , contribute very little to the expression of $H(p_1, p_2, \dots, p_k)$ in comparison to other terms. Therefore, in calculating the entropy all the low-probability outcomes can often be disregarded without risk of any significant error (cf. the text in small print on p. 59). Conversely, in a range between $p = 0.2$ and $p = 0.6$, where $-p \log p$ assumes the greatest value, it changes comparatively evenly; hence in this range even a fairly significant variation in the probabilities p_i has a comparatively small effect on the value of the entropy. We also note that from the continuity of the graph of the function $-p \log p$ it follows that the entropy $H(\alpha)$ depends *continuously* upon the probabilities of individual outcomes of an experiment α , i.e., that, for very small variations of these probabilities, the entropy also varies very little.

Problem 16. *There are two urns, each containing 20 balls, there being 10 white, 5 black and 5 red balls in the first and 8 white, 8 black and 4 red in the second. One ball at a time is drawn from each urn. The outcome of which of these two experiments should be regarded as more uncertain?*

The probability table for the corresponding experiments (we denote them by α_1 and α_2) has the form :

(i) *Experiment α_1 (draw of ball from the first urn) :*

Colour of the ball	white	black	red
Probabilities	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$

(ii) *Experiment α_2 (draw of ball from the second urn) :*

Colour of the ball	white	black	red
Probabilities	$\frac{2}{5}$	$\frac{2}{5}$	$\frac{1}{5}$

The entropy of the first experiment is given by

$$\begin{aligned} H(\alpha_1) &= -\frac{1}{2} \log \frac{1}{2} - \frac{1}{4} \log \frac{1}{4} - \frac{1}{4} \log \frac{1}{4} \\ &= \frac{1}{2} \times 1 + \frac{1}{2} \times 2 = 1.5 \text{ bits,} \end{aligned}$$

but the entropy of the second is somewhat greater, given by

$$\begin{aligned} H(\alpha_2) &= -\frac{2}{5} \log \frac{2}{5} - \frac{2}{5} \log \frac{2}{5} - \frac{1}{5} \log \frac{1}{5} \\ &\approx \frac{4}{5} \times 1.32 + \frac{1}{5} \times 2.32 \approx 1.52 \text{ bits,} \end{aligned}$$

Hence, if we evaluate (as we agreed in order to accomplish this) the amount of uncertainty of the outcome of an experiment from its entropy, then we have to regard the outcome of the second experiment to be more uncertain than that of the first.

Problem 17. *Suppose it to be known from several years of weather observation that for a certain locality the probability that 15 June will be or will not be a rainy day equals 0.4 or 0.6, respectively. Further, assume that for the same locality the probability that on 15 November there will be rain equals 0.65, that there will be snowfall equals 0.15 and the probability that on this day there will be no precipitation equals 0.2. If, of all the weather characteristics, the question of the presence and nature of precipitation alone is of interest, then on which of the two days, enumerated above, should the weather be regarded to be more uncertain in the locality under consideration?*

According to what is understood here by the term 'weather', experiments α_1 and α_2 , which consist of determining the weather that will prevail on 15 June and 15 November, are characterized by the following probability tables:

(i) *Experiment α_1 :*

Outcomes of experiment	Rain	Absence of precipitation
Probabilities	0.4	0.6

(ii) *Experiment α_2 :*

Outcomes of experiment	Rain	Snowfall	Absence of precipitation
Probabilities	0.65	0.15	0.2

Hence, the entropies of our two experiments are given by

$$H(\alpha_1) = -0.4 \log 0.4 - 0.6 \log 0.6 \approx 0.97 \text{ bits},$$

and

$$\begin{aligned} H(\alpha_2) &= -0.65 \log 0.65 - 0.15 \log 0.15 - 0.2 \log 0.2 \\ &\approx 1.28 \text{ bits} > H(\alpha_1). \end{aligned}$$

Consequently, in the locality in question, the weather should be considered to be more uncertain on 15 November than on 15 June.

The result obtained obviously depends substantially on how the term 'weather' is interpreted; without making precisely explicit what is implied by it, our problem in general has no meaning. In particular, if we are only interested in

whether there will be or will not be precipitation on a given day, the two outcomes 'rain' and 'snowfall' of experiment α_2 ought to be combined. For this, instead of α_2 , we have the experiment α'_2 whose entropy is defined by

$$H(\alpha'_2) = -0.8 \log 0.8 - 0.2 \log 0.2 \approx 0.72 < H(\alpha_1).$$

Hence, with such an interpretation of weather, it is necessary to regard the weather to be *less uncertain* on 15 November than on 15 June. If, however, not only the precipitation but, say, the atmospheric temperature is also of concern, then the solution of the problem becomes more complicated and demands that we produce additional data on the temperature distribution in the given locality on 15 June and 15 November.

The arguments developed in the solution of Problem 17 are of interest for an estimate of the quality of weather prediction by some method (the same situation holds for every other forecast). In fact, in an estimate of the quality of prediction, one should not take note of its accuracy alone (i.e., the percentage of cases in which the forecast is fulfilled); otherwise, it would lead to an over-estimation of every forecast that has great chance of being found correct (e.g., of a forecast, say, that there will be no snow in Moscow on 1 June, which is obviously of no importance). For a comparison of the quality of different forecasts, we ought to take note of not only their accuracy but also of the difficulty in making a good forecast, which can be characterized by the amount of uncertainty in the corresponding experiment. We shall again turn to this question later (see Problem 21 in Sec. 3 of this chapter).

Historically, the first steps in the formulation of the concept of entropy were taken as early as 1928 by Hartley[†], the American communication engineer. He suggested to characterize the amount of uncertainty of an experiment with k different outcomes by the number $\log k$. He was, of course, well aware that this measure of uncertainty is quite convenient only in some practical problems, while in many cases it will be quite futile (and even elusive). This is due to the fact that it completely ignores the distinctions among the natures of the occurring outcomes (a most improbable outcome is given here the same importance as a highly likely outcome). However, he held that the distinctions among probable and unlikely outcomes are determined in the first place by 'psychological factors' and, therefore, should be taken into account only by psychologists and not be considered by communication engineers and mathematicians.

The fallibility of Hartley's viewpoint was shown by C. Shannon in 1948. He introduced the quantity

$$H(x) = -p(A_1) \log p(A_1) - p(A_2) \log p(A_2) - \dots - p(A_k) \log p(A_k),$$

[†]R. V. L. Hartley (1928). Transmission of information, *Bell System Tech. J.* 7(3), 535-63.

as a measure of uncertainty of an experiment α with A_1, A_2, \dots, A_k possible outcomes, where $p(A_1), p(A_2), \dots, p(A_k)$ are the probabilities of individual outcomes; he also named this quantity 'entropy'. In other words, Shannon assigns the uncertainty $-\log p(A_i)$ to an outcome A_i of the event α (in the case of k equally likely outcomes with probability $p = 1/k$, it leads to old Hartley's suggestion to take the number $\log k = -\log p$ as a measure of uncertainty). Furthermore, as a measure of the uncertainty of the entire experiment α , we take the *mean value* of the uncertainty of individual outcomes, i.e., the mean value of a random variable taking the values $-\log p(A_1), -\log p(A_2), \dots, -\log p(A_k)$ with probabilities $p(A_1), p(A_2), \dots, p(A_k)$ [by the definition derived on p. 6 this mean value is precisely equal to $H(\alpha)$]. Thus, the perplexing 'psychological factors' introduced by Hartley are here taken into account by using the concept of probability having a purely mathematical (or more accurately, a statistical) character.

The use of the quantity $H(\alpha)$ as a measure of uncertainty of the experiment α is found very convenient for a large variety of purposes; in the following, our main objective will be to make this situation transparent. It should, however, be borne in mind that the Shannon measure, as also Hartley's measure, cannot lay claim to take into account all factors, determining the 'uncertainty of an experiment' in every sense in which it may be encountered in real life. Thus, for example, $H(\alpha)$ depends only on the probabilities $p(A_1), p(A_2), \dots, p(A_k)$ of the various outcomes of the experiment but in no way depends on what these outcomes are, whether they are in a certain sense 'close' to or quite 'remote' from each other. Hence, our 'amount of uncertainty' will be the same for two random variables characterized by the following probability tables :

<i>Values</i>	0.9	1	1.1
<i>Probabilities</i>	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

and

<i>Values</i>	- 200	1	1000
<i>Probabilities</i>	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$

or for two methods of treatment of a patient, of which one results in complete recovery of 90 out of 100 cases and appreciable improvement in the condition of the patient in the remaining 10 cases, and the second also achieves complete success in 90 out of 100 cases but then in the remaining 10 cases it is concluded by lethal outcome. The vital distinction between the two experiments in these cases has to be evaluated by other characteristics, different from Shannon's entropy.

The peculiarity of the entropy $H(\alpha)$ indicated, as also a series of other singularities of this quantity, stem naturally from the fact that the concept of entropy

first arose in attempting to solve particularly some problems intimately related to the transmission of information through communication lines and hence it is very suitable for precisely such uses. The situation is that for determining the time required for the transmission of some communication or the cost of such transmission, the *specific content* of the communication itself is altogether immaterial! This is manifested in the entropy $H(\alpha)$ being independent of the values A_1, A_2, \dots, A_k of the outcomes of an experiment. On the other hand, the *probabilities* of individual communication are of great importance for communication theory; we shall further elaborate on this in Chapter 4. Another property of great importance is that in the operation of communication lines a crucial role is played by *statistical* regularities, since through these lines there is always transmitted a large amount of information of various kinds. Hence, the measure of uncertainty used in the solution of problems related to communication engineering must first be adapted to the evaluation of the amount of uncertainty of intricate 'compound experiments' consisting of a long series of trials following one after another.

Let us also note that, from the viewpoint of an investigator studying the amount of uncertainty of such compound experiments, the difference between the treatments due to Shannon and Hartley is not found to be as striking as it might appear at the start. In fact, even if we look from Hartley's standpoint, it is impossible to ignore completely the probabilities of the occurrence of outcomes, otherwise we could arbitrarily increase the number k of outcomes of our experiment, adding to really possible outcomes any number of fictitious outcomes of probability zero. Hence, in the calculation of the measure of uncertainty of an experiment, according to Hartley, we should certainly reject all 'impossible' outcomes of zero probability. However, in addition, it is hardly worthwhile to take account of 'practically impossible' outcomes having negligibly small probability of occurrence. We now replace the experiment α with k distinct outcomes by another experiment α_N made up of a number of repetitions N (under identical conditions) of α . The number of distinct outcomes of α_N is k^N ; these k^N outcomes are obtained by combining the k possible outcomes of the first performance of α with the k possible outcomes of the second performance, \dots , k outcomes of the N th performance of α . Hence, the amount of uncertainty of experiment α_N , by Hartley's measure is $\log k^N = N \log k$, which again leads to the expression $\log k$ for the amount of uncertainty of α (because it is natural to consider that the amount of uncertainty of an event which consists of a number N of repetitions of α , must be N times greater than the amount of uncertainty of α ; cf. a similar argument on p. 45).

However, so far we have said nothing about the *probabilities* of our k^N outcomes of the event α_N . It is plain that if k outcomes of α are equally likely, then all k^N outcomes of experiment α_N are equally likely also, since here none of these k^N events is distinguished by anything from the rest. If, however, k outcomes of α have the different probabilities $p(A_1), p(A_2), \dots, p(A_k)$, then

$k^N = 2^{N \log k}$ outcomes of the compound event α_N have also the different probabilities. It is found that, for large values of N , most of these $2^{N \log k}$ outcomes will have such *negligible* probability that even the sum of the probabilities of *all* such low probability outcomes is very small. As regards the remaining (more probable) outcomes of the experiment α_N , the probabilities of all these outcomes for large N are almost indistinguishable from each other. Speaking more precisely, it can be shown *that for sufficiently large N we can always discard some (as a rule a quite large !) portion of the outcomes of an experiment α_N , so that the total probability of all the excluded outcomes is less than any quite small number chosen beforehand* (say, less than 0.01, or 0.001, or 0.000001; the only requirement here is that the smaller we choose this number to be, the greater the number N should be) *and all the remaining outcomes of the experiment α_N have practically the same probability*. It is highly important in this case that the *number of outcomes of the experiment α_N , left over after such rejection, is found to be of order $2^{N H(\alpha)}$, where $H(\alpha) = -p(A_1) \log p(A_1) - \dots - p(A_k) \log p(A_k)$ is the entropy of experiment α* .[†] Hence, it is clear that even from Hartley's viewpoint it is natural to take the number $\log 2^{N \cdot H(\alpha)} = N \times H(\alpha)$ as a measure of the uncertainty of the experiment α_N (because the outcomes, whose probability sum is negligible, are naturally discarded); in addition, for the amount of uncertainty of the initial experiment α , we obtain the value $N \times H(\alpha)/N = H(\alpha)$. Thus, it is seen that the treatment of Shannon differs from that of Hartley primarily in building up a long chain of repeated realizations of one and the same experiment α ; the consideration of such a chain is typical of a probabilistic (statistical) approach.

The statement, set above in italic letters, brings out the statistical meaning of the concept of entropy; it lies at the very root of most of the engineering applications of this concept. However, a proof of this statement is not quite straightforward; we defer it (and also a somewhat more precise formulation of the statement itself) to the last section, which is directly devoted to the applications of the concept of entropy to the theory of transmission of information.

The real value of the concept of entropy stems primarily from the fact that the 'amount of uncertainty' of an experiment expressed by it is found in many cases to be that particular characteristic, which has a role to play in diverse processes of the transmission and storage of various types of information encountered in nature and engineering. Later, we shall give a more elaborate exposition of some engineering applications of the concept of entropy; here, however, we shall present only one example of an entirely different variety.

One of the basic problems dealt with in experimental psychology is a study of *psychic reactions*, i.e., the response of an organism to some stimulation or action.

[†]Here, it follows in particular that if only all outcomes of the experiment α are not equally likely and, consequently, $H(\alpha) < \log k$, then the number of excluded outcomes form a dominating part of all the outcomes of α_N (because the ratio $2^{N \cdot H(\alpha)} : k^N = 2^{N \cdot H(\alpha)} : 2^{N \cdot \log k} = 2^{-N \cdot (\log k - H(\alpha))}$ is quite small for large N).

In addition, these reactions are classified into a *simple reaction*, some definite response to some assigned signal, and a *complex reaction*, the most important of which is the *reaction of choice*, in which different signals evoke different responses. It is known that the time necessary for a simple reaction of a person (i.e., the time interval between the stimulus and the reaction) does not depend usually on the nature of the stimulating signal (for mature people, its minimum value is close to 0.1 sec). A considerably more complicated problem is to ascertain the time necessary for a complex reaction. This depends substantially on the conditions of the experiment and primarily on the 'amount of complexity' of the reaction. As early as the 1880's psychologists had explained that the average rate at which a person can react to a sequence of random consecutive signals of k different kinds (provided that to each kind of signal he must react differently) decreases monotonically with increasing k . In order to verify this fact, a large number of experiments were carried out to determine the average time necessary for a chosen reaction, and these almost always yielded roughly the same result. The most usual setting of such experiments is a board before the subject on which one of k lights is flashed or one of k digits appears at definite intervals of time, and depending on the number of signals that appear he is to press one of the k buttons on which he had his fingers beforehand or utter one of the preassigned k words. A special device records the time transpiring between the appearance of the signal and the reaction of the subject to it; the dependence of the mean reaction time T obtained on the number k is also studied.

It is natural to consider the mean reaction time as a definite 'measure of uncertainty' of the expected signal: the greater the uncertainty in the occurrence of the event, the greater is the time required to ascertain precisely what signal was delivered. The existing experimental data show that the *mean reaction time increases with the increase of the number k of different kinds of signals roughly as $\log k$* , i.e., as the *Shannon entropy $H(\alpha)$ of an experiment α , consisting of sending the signals* (in all the experiments with which we are concerned here, the probability of signals of different kinds is always the same). For example, in Fig. 10 (taken from R. Hyman [47]) the circles show the data of eight experiments which were carried out to determine the average time required by the subject, to indicate which of k lights was flashed. The number k ranged in these experiments from 1 to 8. The mean reaction time was determined from a large number of series of flashes, in each of which the flash frequency of all lights was identical, and the subject was already especially trained in similar experiments. In Fig. 10, the ordinate gives the mean reaction time and the abscissa the quantity $\log k$; in addition, as is seen, all eight circles are laid sufficiently precisely along a single straight line.

On the basis of this data, it is natural to surmise that the *mean reaction time in all cases is determined by the entropy of an experiment α consisting of sending the signals*. This, in turn, implies that the decrease in the degree of uncertainty of an experiment caused by replacing equally probable signals by not equally

probable signals must produce the same reduction of the mean reaction time as occurs when the number of different kinds of signals used is decreased, leading to the same reduction in the entropy $H(\alpha)$. This statement admits direct experimental verification, which substantiates it completely. Thus, in the same Fig. 10, the squares plot the results of eight other experiments (carried out with the

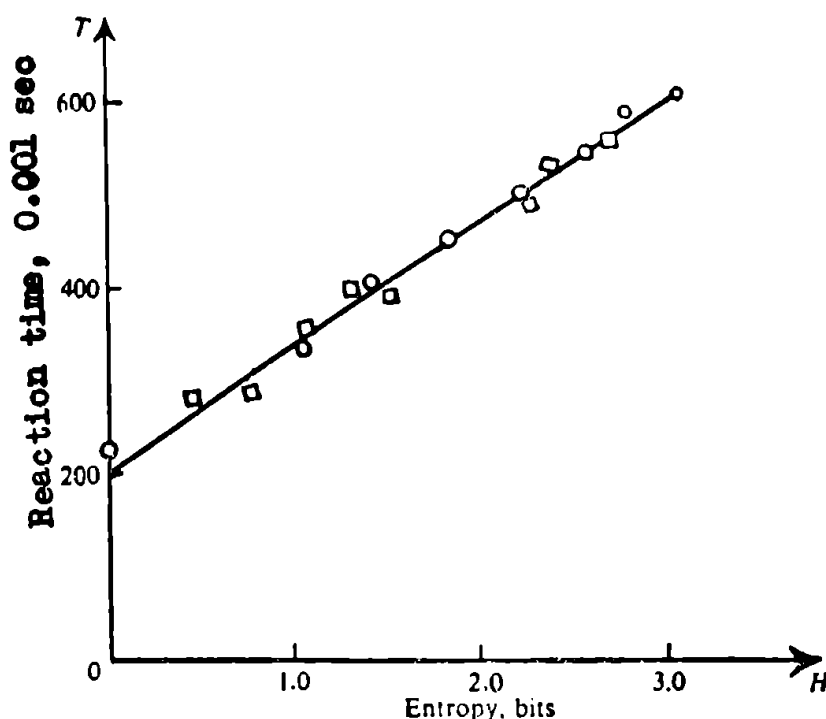


Fig. 10.

same subject as earlier) where k lights (with k equal to 2, 4, 6 or 8) were flashed with different relative frequencies $p(A_1)$, $p(A_2)$, \dots , $p(A_k)$, and the subject was trained beforehand for some time to a series of flashes at such frequencies. Here again, the mean reaction time T is plotted on the ordinate and the entropy $H(\alpha) = -p(A_1) \log p(A_1) - p(A_2) \log p(A_2) - \dots - p(A_k) \log p(A_k)$ on the abscissa; it is found here that the squares are arranged quite accurately along the same straight line along which circles lie. It is thus seen that the entropy $H(\alpha)$ is indeed precisely that measure of the uncertainty of the outcome of an experiment which determines in a decisive manner the mean time required for a specific reaction to the advent of a signal.

The reason for variation in the mean reaction time with variation in the relative frequencies of different signals is implicit in the fact that the subject reacts appropriately more rapidly to more frequently appearing (i.e., more familiar to him) signals but, on the other hand, more slowly to infrequent signals which are not expected by him. Obviously, these factors bear a psychological character. Nevertheless, it is seen that they can also be characterized quantitatively by the

value of entropy $H(\alpha)$ of an experiment α , despite Hartley's misgivings that no 'psychological factors' (which, however, according to his understanding had a considerably more vague relation with psychology than in the present example) can be quantitatively estimated.

At the end of this section we deduce some data, characterizing the insignificant value of numerous low-probability outcomes in determining the entropy of an experiment with many outcomes.

We consider an experiment in which we are to select at random an English word consisting of four letters from a printed text. We can use here the data contained in the well-known 'Thorndike dictionary' [167], which catalogues the frequencies of 20,000 most common English words, obtained by a statistical analysis of quite voluminous and varied English texts. This dictionary contains altogether 1550 four-letter words; accordingly, we can consider that our experiment α has 1550 different outcomes. We now calculate the entropy

$$H(\alpha) = -p(A_1) \log p(A_1) - p(A_2) \log p(A_2) - \dots - p(A_{1550}) \log p(A_{1550})$$

of this experiment, taking the probability $p(A_i)$ of each outcome to be equal to the frequency n_i/N of the corresponding word; here n_i is the number of the repetitions of this word, catalogued in Thorndike's dictionary, and $N = n_1 + n_2 + \dots + n_{1550}$. It is found that this entropy is close to 8.14 bits†. We shall now discard all words for which $n_i < 150$; by doing so, there remain only 865 four-letter words, i.e., slightly more than 50 per cent of the original number (to be precise, 55.8 per cent). At the same time, a part of the sum $H(\alpha)$, corresponding to these 865 words, is equal to roughly 8 bits, i.e., forms more than 98 per cent of the entire quantity $H(\alpha)$. We now reject all words for which $n_i < 750$; by doing so, we are left with 395 words, i.e., in all about one-fourth (25.5 per cent) of the original number. However, to these 395 words there corresponds a part of the sum $H(\alpha)$, greater than 7.47 bits, i.e., constituting over 92 per cent of the entire quantity $H(\alpha)$. If we next exclude all words with $n_i < 1550$, then we are left with only 214 words (13.8 per cent of the original number); however, to these 214 outcomes there corresponds a part of the sum $H(\alpha)$, close to 6.88 bits, i.e., comprising about 85 per cent of its original value. Finally, if we discard all words with $n_i < 3150$, then altogether 119 four-letter words are left (7.7 per cent of the original number); however, to this 7.7 per cent of outcomes there corresponds roughly 78 per cent of the sum $H(\alpha)$ (this part of the sum $H(\alpha)$ exceeds 6.44 bits).

2.2. The entropy of compound events. Conditional entropy

Let α and β be two *independent* experiments with the following probability :

(i) *Experiment α :*

Outcomes of experiment	A_1	A_2	\dots	A_k
Probabilities	$p(A_1)$	$p(A_2)$	\dots	$p(A_k)$

†This value, as also all the accompanying numerical data, are taken from [19].

of l members in the expression for $H(\alpha\beta)$ are given by

$$\begin{aligned} & -p(A_2) \log p(A_2) + p(A_2) H(\beta), \\ & \dots\dots\dots \\ & -p(A_k) \log p(A_k) + p(A_k) H(\beta) \end{aligned}$$

and, hence,

$$\begin{aligned} H(\alpha\beta) = & -p(A_1) \log p(A_1) - p(A_2) \log p(A_2) - \dots - p(A_k) \log p(A_k) \\ & + (p(A_1) + p(A_2) + \dots + p(A_k)) H(\beta) = H(\alpha) + H(\beta) \end{aligned}$$

(since, also $p(A_1) + p(A_2) + \dots + p(A_k) = 1$).

We now assume that experiments α and β are *not independent* (for example, that α and β are the successive draws of two balls from one urn; see p. 20). In this more general case, we cannot expect that the entropy of the compound experiment $\alpha\beta$ is the sum of the entropies of α and β . In fact, a case can be conceived here such that the result of the second experiment is completely determined by the result of the first (for example, if the experiments α and β consist of the successive draws of two balls from an urn, containing in all two balls of different colours). Thus, after realization of the experiment α , the experiment β *completely loses* its uncertainty; hence, here it is natural to assume that the entropy (the measure of the amount of uncertainty) of the compound experiment $\alpha\beta$ equals the entropy of the single experiment α but not the sum of the entropies of α and β (in the following, we shall be convinced that it is indeed so). We shall attempt to make explicit the expression by defining the entropy of $\alpha\beta$ in a general case.

We reiterate the conclusion of the formula for the entropy $H(\alpha\beta)$ of $\alpha\beta$, *without* the supposition of α and β being independent. Obviously, as before, we have

$$\begin{aligned} H(\alpha\beta) = & -p(A_1B_1) \log p(A_1B_1) - p(A_1B_2) \log p(A_1B_2) - \dots \\ & - p(A_1B_l) \log p(A_1B_l) - p(A_2B_1) \log p(A_2B_1) \\ & - p(A_2B_2) \log p(A_2B_2) - \dots - p(A_2B_l) \log p(A_2B_l) \\ & \dots\dots\dots \\ & - p(A_kB_1) \log p(A_kB_1) - p(A_kB_2) \log p(A_kB_2) - \dots \\ & - p(A_kB_l) \log p(A_kB_l), \end{aligned}$$

where by A_1, A_2, \dots, A_k and B_1, B_2, \dots, B_l we again denote, respectively, the outcomes of α and β . However, here it is impossible to replace the probabilities $p(A_1B_1), p(A_1B_2) \dots$ simply by the products of corresponding probabilities. In fact, $p(A_1B_1)$ is now not equal to $p(A_1) p(B_1)$, but it is equal to $p(A_1) p_{A_1}(B_1)$, where $p_{A_1}(B_1)$ is the *conditional probability* of the event B_1 given A_1 (see Sec. 3, Chap. 1). This circumstance is prominently manifested in the following reasoning.

occurring with positive probabilities, then $H_{\alpha}(\beta) = 0$ if and only if $H_{A_1}(\beta) = H_{A_2}(\beta) = \dots = H_{A_k}(\beta) = 0$, i.e., if and only if for every outcome of the experiment α , the result of the experiment β stands completely determined (trivially, this condition is satisfied if the experiment β is not indeterminate from the very outset). In such a case, we have

$$H(\alpha\beta) = H(\alpha)$$

(see p. 61). If, however, the experiments α and β are independent, then

$$H_{A_1}(\beta) = H_{A_2}(\beta) = \dots = H_{A_k}(\beta) = H(\beta)$$

and, hence,

$$H_{\alpha}(\beta) = H(\beta).$$

In this case, the formula $H(\alpha\beta) = H(\alpha) + H_{\alpha}(\beta)$ carries over into a simpler one: $H(\alpha\beta) = H(\alpha) + H(\beta)$ (see p. 60).

It is quite essential that in all cases the *conditional entropy* $H_{\alpha}(\beta)$ lies between 0 and the (unconditional) entropy $H(\beta)$ of β , which is sometimes called the *marginal entropy* of β :

$$0 \leq H_{\alpha}(\beta) \leq H(\beta),$$

implying that *the conditional entropy can never be greater than the unconditional one*. Thus, the two cases, namely, when an outcome of the experiment β is completely determined by an outcome of α and when α and β are independent, are two extreme cases.

The statement that $0 \leq H_{\alpha}(\beta) \leq H(\beta)$ is also in good agreement with the interpretation of entropy as a measure of uncertainty: it is completely obvious that the previous realization of the experiment α can only decrease the amount of uncertainty of β or, in the extreme case (say, in the case when α and β are independent), does not change this amount of uncertainty, but in no case it can increase it†.

A complete proof of the statement made (including also a proof of the fact that $H_{\alpha}(\beta) = H(\beta)$ if and only if the experiments α and β are independent) is given in Appendix I at the end of the book; here we shall only demonstrate it by an example in which an experiment α has two *equally probable* outcomes,

†To avoid possible fallacy, we note that the conditional entropy $H_{A_1}(\beta)$ can be both smaller and greater than the unconditional entropy $H(\beta)$ (see, for example, Problems 18 and 19 below). This is related to the fact that a change in the probability table of the experiment β , postulated by the circumstances that the other experiment α had a definite outcome A_1 , can be sufficiently arbitrary (see pp. 21-22).

A_1 and A_2 . In this case

$$H_{\alpha}(\beta) = p(A_1) H_{A_1}(\beta) + p(A_2) H_{A_2}(\beta) = \frac{1}{2} H_{A_1}(\beta) + \frac{1}{2} H_{A_2}(\beta).$$

Thus, our problem reduces to establishing that the inequality

$$\frac{1}{2} H_{A_1}(\beta) + \frac{1}{2} H_{A_2}(\beta) \leq H(\beta)$$

holds. In other words, it is required to show that

$$\begin{aligned} & \frac{1}{2} [-p_{A_1}(B_1) \log p_{A_1}(B_1) - p_{A_1}(B_2) \log p_{A_1}(B_2) - \dots - p_{A_1}(B_l) \log p_{A_1}(B_l)] \\ & + \frac{1}{2} [-p_{A_2}(B_1) \log p_{A_2}(B_1) - p_{A_2}(B_2) \log p_{A_2}(B_2) - \dots \\ & - p_{A_2}(B_l) \log p_{A_2}(B_l)] \\ & \leq -p(B_1) \log p(B_1) - p(B_2) \log p(B_2) - \dots - p(B_l) \log p(B_l), \end{aligned}$$

where B_1, B_2, \dots, B_l are outcomes of the experiment β . We again consider the graph of the function $F(x) = -x \log x$, and suppose that

$$OA = p_{A_1}(B_1), \quad OB = p_{A_2}(B_1)$$

in Fig. 11. Then, the segments AM and BN are of lengths $-p_{A_1}(B_1) \log p_{A_1}(B_1)$

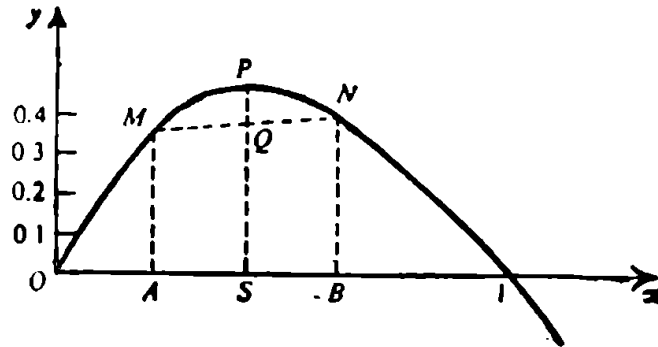


Fig. 11.

and $-p_{A_2}(B_1) \log p_{A_2}(B_1)$, respectively. The sum $-\frac{1}{2} p_{A_1}(B_1) \log p_{A_1}(B_1) - \frac{1}{2} p_{A_2}(B_1) \log p_{A_2}(B_1)$ is equal to the middle line SQ of the trapezium $ABNM$. On the other hand, the segment $SP > SQ$ is equal to $-p(B_1) \log p(B_1)$, since

$$OS = \frac{1}{2} OA + \frac{1}{2} OB = p(A_1) p_{A_1}(B_1) + p(A_2) p_{A_2}(B_1) = p(B_1)$$

(see the equation of total probability on p. 23). Consequently,

$$-\frac{1}{2} p_{A_1}(B_1) \log p_{A_1}(B_1) - \frac{1}{2} p_{A_2}(B_1) \log p_{A_2}(B_1) \leq -p(B_1) \log p(B_1).$$

The experiment α also has two outcomes : A_1 (positive reaction) and A_2 (negative reaction). The probabilities of these two outcomes are given by

$$p(A_1) = 0.51 \quad \text{and} \quad p(A_2) = 0.49.$$

(This is because the outcome A_1 occurs in one-half of those cases in which β has an outcome B_1 and in all cases in which β has an outcome B_2 , but the outcome A_2 is realized only in one-half of the cases in which β has an outcome B_1 .) Moreover, if α had A_1 outcomes, then the conditional probabilities of the outcomes of β are given by

$$p_{A_1}(B_1) = \frac{49}{51} \quad \text{and} \quad p_{A_1}(B_2) = \frac{2}{51}$$

(because out of 51 cases in which the reaction is positive, a person is found healthy in 49 cases, and sick in two cases); hence the conditional entropy $H_{A_1}(\beta)$ will be appreciably greater than the unconditional entropy $H(\beta)$:

$$H_{A_1}(\beta) = -\frac{49}{51} \log \frac{49}{51} - \frac{2}{51} \log \frac{2}{51} \approx 0.24 \text{ bits.}$$

On the other hand, if the experiment α has an outcome A_2 , then we can state *with certainty* that the experiment β had an outcome B_1 (the person is healthy); consequently,

$$H_{A_2}(\beta) = 0.$$

Thus, the mean conditional entropy of β given that α is realized is less than the unconditional entropy $H(\beta)$:

$$H_\alpha(\beta) = 0.51 \times H_{A_1}(\beta) + 0.49 \times H_{A_2}(\beta) \approx 0.51 \times 0.24 \approx 0.12 \text{ bits.}$$

In other words, the realization of α decreases the amount of uncertainty of β by roughly 0.02 bits.

Problem 19. Suppose that the experiments α and β consist of drawing successively two balls from an urn, containing m black and $n - m$ white balls (α is the draw of the first ball, β the draw of the second ball). Determine the entropies $H(\alpha)$ and $H(\beta)$ and the conditional entropies $H_\beta(\alpha)$ and $H_\alpha(\beta)$ of α and β , respectively. Solve this problem also subject to the condition that experiment α consists of drawing k balls from the urn and experiment β is the succeeding draw of one more ball.

We start with the case when α consists of the draw of one ball. We suppose that A_1 and A_2 (resp. B_1 and B_2) represent the appearance of a black and a white ball in the first (resp. second) draw. When nothing is known about the outcomes of either the first or the second experiment, we can expect the realization of these events with the following probabilities :

(i) Experiment α :	Outcomes of experiment	A_1	A_2
	Probabilities	$\frac{m}{n}$	$\frac{n-m}{n}$
(ii) Experiment β :	Outcomes of experiment	B_1	B_2
	Probabilities	$\frac{m}{n}$	$\frac{n-m}{n}$

Thus, both these experiments have the same entropy :

$$H(\alpha) = H(\beta) = -\frac{m}{n} \log \frac{m}{n} - \frac{n-m}{n} \log \frac{n-m}{n}.$$

If the outcomes of the experiment α be known to us, then, the probabilities of the individual outcomes of the experiment β will have different values. To be exact (see above p. 20) :

$$\begin{aligned} p_{A_1}(B_1) &= \frac{m-1}{n-1}, & p_{A_1}(B_2) &= \frac{n-m}{n-1}; \\ p_{A_2}(B_1) &= \frac{m}{n-1}, & p_{A_2}(B_2) &= \frac{n-m-1}{n-1}. \end{aligned}$$

Hence it follows that

$$\begin{aligned} H_{A_1}(\beta) &= -\frac{m-1}{n-1} \log \frac{m-1}{n-1} - \frac{n-m}{n-1} \log \frac{n-m}{n-1}, \\ H_{A_2}(\beta) &= -\frac{m}{n-1} \log \frac{m}{n-1} - \frac{n-m-1}{n-1} \log \frac{n-m-1}{n-1}. \end{aligned}$$

Further, if $m < n-m$, then

$$H_{A_1}(\beta) < H(\beta), \quad H_{A_2}(\beta) > H(\beta)$$

(because the uncertainty of an experiment, consisting of the draw of a ball from an urn with m black and $m_1 = n-m$ white balls, increases as the ratio m/m_1 approaches unity). Finally, we have

$$\begin{aligned} H_{\alpha}(\beta) &= p(A_1)H_{A_1}(\beta) + p(A_2)H_{A_2}(\beta) \\ &= \frac{m}{n} \left[-\frac{m-1}{n-1} \log \frac{m-1}{n-1} - \frac{n-m}{n-1} \log \frac{n-m}{n-1} \right] \\ &\quad + \frac{n-m}{n} \left[-\frac{m}{n-1} \log \frac{m}{n-1} - \frac{n-m-1}{n-1} \log \frac{n-m-1}{n-1} \right] \end{aligned}$$

(in all cases $H_{\alpha}(\beta) < H(\beta)$) and

$$H_{\beta}(\alpha) = H_{\alpha}(\beta) + \{H(\alpha) - H(\beta)\} = H_{\alpha}(\beta),$$

We now pass on to the more general problem set forth in the second hypothesis. We denote by α_k an experiment α consisting of the draw of k balls from an urn, and assume that k does not exceed the numbers m and $n - m$. In such a case, α_k can have $k + 1$ different outcomes corresponding to the fact that among the subject balls there are $0, 1, 2, \dots, k$ black balls; we denote these outcomes by $A_0, A_1, A_2, \dots, A_k$. The probability $p(A_i)$ of the outcome A_i ($i = 0, 1, \dots, k$) equals the ratio $\binom{m}{i} \binom{n-m}{k-i} / \binom{n}{k}$. In fact, the total number of equally probable outcomes of the experiment α_k equals $\binom{n}{k}$ (the number of all possible groups of k balls which can be composed of the available n balls), and of them the outcomes $\binom{m}{i} \binom{n-m}{k-i}$ are favourable to the outcome A_i (since out of the m available balls i black balls can be selected in $\binom{m}{i}$ ways, and the remaining $k - i$ white balls in $\binom{n-m}{k-i}$ ways). This implies that the entropy of α_k is

$$\begin{aligned}
 H(\alpha_k) = & - \frac{\binom{n-m}{k}}{\binom{n}{k}} \log \frac{\binom{n-m}{k}}{\binom{n}{k}} - \frac{\binom{m}{1} \binom{n-m}{k-1}}{\binom{n}{k}} \log \frac{\binom{m}{1} \binom{n-m}{k-1}}{\binom{n}{k}} \\
 & - \frac{\binom{m}{2} \binom{n-m}{k-2}}{\binom{n}{k}} \log \frac{\binom{m}{2} \binom{n-m}{k-2}}{\binom{n}{k}} - \dots \\
 & - \frac{\binom{m}{k-1} \binom{n-m}{1}}{\binom{n}{k}} \log \frac{\binom{m}{k-1} \binom{n-m}{1}}{\binom{n}{k}} \\
 & - \frac{\binom{m}{k}}{\binom{n}{k}} \log \frac{\binom{m}{k}}{\binom{n}{k}} .
 \end{aligned}$$

The experiment β has two outcomes, B_1 (the draw of a black ball) and B_2 (the draw of a white ball). The probability of these two outcomes is, respectively, equal to m/n and $(n - m)/n$. The entropy of β , as before, is

$$H(\beta) = - \frac{m}{n} \log \frac{m}{n} - \frac{n-m}{n} \log \frac{n-m}{n} .$$

Now suppose it to be known that the outcome A_i of the experiment α_k has occurred. This means that, after the realization of this experiment, $m - i$ black and $n - m - k + i$ white balls are left in the urn. In conformity with this

$$p_{A_i}(B_1) = \frac{m-i}{n-k}, \quad p_{A_i}(B_2) = \frac{n-m-k+i}{n-k}$$

and

$$H_{A_i}(\beta) = -\frac{m-i}{n-k} \log \frac{m-i}{n-k} - \frac{n-m-k+i}{n-k} \log \frac{n-m-k+i}{n-k}.$$

To calculate $H_{\alpha_k}(\beta)$ it remains only to make use of the formula

$$\begin{aligned} H_{\alpha_k}(\beta) &= \frac{\binom{n-m}{k}}{\binom{n}{k}} H_{A_0}(\beta) + \frac{\binom{m}{1} \binom{n-m}{k-1}}{\binom{n}{k}} H_{A_1}(\beta) + \dots \\ &\quad + \frac{\binom{m}{k}}{\binom{n}{k}} H_{A_k}(\beta). \end{aligned}$$

Finally, the conditional entropy $H_{\beta}(\alpha_k)$ is defined by

$$H_{\beta}(\alpha_k) = H_{\alpha_k}(\beta) + H(\alpha_k) - H(\beta).$$

The case when k is greater than either the number m or $n-m$ or even both can be treated similarly. We shall not analyze here all the possibilities open to us but confine ourselves only to a few observations.

(a) Suppose that $k = n-1$. The experiment α_{n-1} has, in all, two outcomes A_1 and A_2 corresponding to the case in which the last ball remaining in the urn is black (white). The probability of these two outcomes equals m/n and $(n-m)/n$ because the choice of $n-1$ balls to be drawn is equivalent to that of the single remaining ball and, consequently, our experiment α_{n-1} does not substantially differ from the experiment α_1 consisting of the draw of exactly one ball from an urn with n balls. Thus, the entropy of α_{n-1} is

$$H(\alpha_{n-1}) = -\frac{m}{n} \log \frac{m}{n} - \frac{n-m}{n} \log \frac{n-m}{n},$$

i.e., it coincides with the entropy of β . As to the conditional entropy $H_{\alpha_{n-1}}(\beta)$, it is obviously 0, because the outcome of α_{n-1} completely predetermines the outcome of β . By analogous reasoning, the conditional entropy $H_{\beta}(\alpha_{n-1})$ is also 0.

(b) Suppose that $k = n-2$. The experiment α_{n-2} has three outcomes, A_0 , A_1 and A_2 , corresponding to the case in which there remain in the urn either two black balls, or a black and a white ball, or two white balls (we assume here that neither of the numbers m and $n-m$ is less than 2). The probabilities of these outcomes are given by

$$p(A_0) = \frac{\binom{m}{2}}{\binom{n}{2}} = \frac{m(m-1)}{n(n-1)}, \quad p(A_1) = \frac{\binom{m}{1} \binom{n-m}{1}}{\binom{n}{2}} = \frac{2m(n-m)}{n(n-1)},$$

$$p(A_2) = \frac{\binom{n-m}{2}}{\binom{n}{2}} = \frac{(n-m)(n-m-1)}{n(n-1)}.$$

In agreement with this, the entropy of α_{n-2} is

$$\begin{aligned} H(\alpha_{n-2}) = & -\frac{m(m-1)}{n(n-1)} \log \frac{m(m-1)}{n(n-1)} - \frac{2m(n-m)}{n(n-1)} \log \frac{2m(n-m)}{n(n-1)} \\ & - \frac{(n-m)(n-m-1)}{n(n-1)} \log \frac{(n-m)(n-m-1)}{n(n-1)}. \end{aligned}$$

The conditional entropy of the experiment β given the realization of a definite outcome of α_{n-2} , is given by†

$$H_{A_0}(\beta) = 0, \quad H_{A_1}(\beta) = 1, \quad H_{A_2}(\beta) = 0,$$

but the conditional entropy of β given the realization of α_{n-2} is

$$H_{\alpha_{n-2}}(\beta) = \frac{2m(n-m)}{n(n-1)}.$$

Finally, the conditional entropy of α_{n-2} given the realization of β is

$$H_{\beta}(\alpha_{n-2}) = H_{\alpha_{n-2}}(\beta) + H(\alpha_{n-2}) - H(\beta).$$

(c) If $m = 1$, then the experiment α_k has just two outcomes A_1 and A_0 corresponding to the cases in which exactly one black ball is found among k balls drawn or among $n - k$ balls remaining in the urn; the probabilities of these outcomes are given by

$$p(A_1) = \frac{k}{n}, \quad p(A_0) = \frac{n-k}{n}.$$

The conditional entropy of the experiment β given that the outcome A_1 of the experiment α_k has occurred is 0 :

$$H_{A_1}(\beta) = 0,$$

(because obviously the outcome A_1 of α_k uniquely determines the outcome of β). The conditional entropy of β given that the outcome A_0 of α_k has occurred is

$$H_{A_0}(\beta) = -\frac{1}{n-k} \log \frac{1}{n-k} - \frac{n-k-1}{n-k} \log \frac{n-k-1}{n-k};$$

it exceeds the (unconditional) entropy of the same experiment β , which is given

†Here $H_{A_1}(\beta) > H(\beta)$, since an experiment β with two outcomes cannot have an entropy exceeding 1 bit.

by

$$H(\beta) = -\frac{1}{n} \log \frac{1}{n} - \frac{n-1}{n} \log \frac{n-1}{n}.$$

(In fact, if among the balls contained in the urn only one ball differs in colour from the rest, then the amount of uncertainty of the experiment, consisting of the draw of one ball, is the smaller, the larger is the total number of balls.) However, the mean conditional entropy of β

$$H_{\alpha_k}(\beta) = \frac{n-k}{n} H_{A_0}(\beta)$$

is less than the (unconditional) entropy $H(\beta)$.

If the pair of experiments α and β are carried out many times one after the other, then the conditional entropy $H_{\alpha}(\beta)$ characterizes that mean amount of uncertainty of the outcome of β which remains after the outcome of the experiment α preceding it is known. In particular, in an experiment on determining

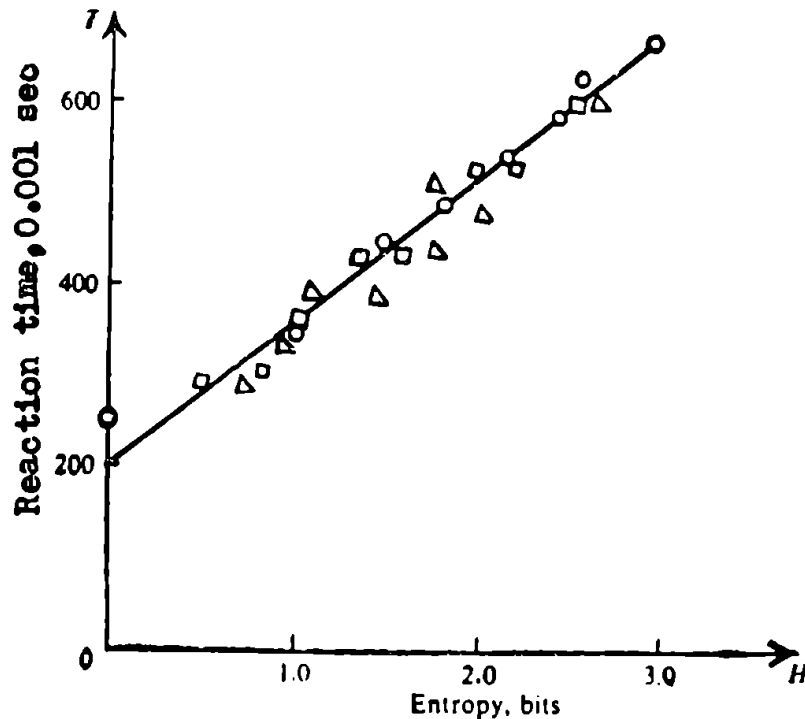


Fig. 12.

average reaction time (see p. 56 and onwards) a whole series of signals is always sent and the subject knows prior to each of them what signals have previously been given to him. Hence, the amount of uncertainty of the signal to be sent here equals the *conditional entropy* of the corresponding experiment given that the outcomes of all previous experiments (i.e., previous sending of signals) are known. In the experiments described on pp. 56-59, the successively sent signals were always selected *independently* of each other; hence, in these experiments, the conditional entropy of α coincided with its unconditional entropy

$H(\alpha)$. If, however, the reaction time is actually determined by the amount of uncertainty of the signal to be sent, measured by its entropy, then from what has been stated above, it necessarily follows that a variation in the amount of uncertainty produced by introducing a dependence among successively sent signals, must have the same influence on the variation of the mean reaction time as a variation in the amount of uncertainty due to a change in the total number of equally probable signals to be used or due to an alteration in the relative frequencies of these signals. The results of the verification of this conclusion are depicted in Fig. 12, taken again from [47]. In this figure are plotted 8 circles and 8 squares, encountered earlier in Fig. 10 and, in addition, 8 triangles corresponding to the results of 8 experiments (performed on the same subjects as previously), in which the subject was required to react differently to flashes of each of the k lights (experiment β ; in the different experiments, k assumed the values 2, 3, 4, 5 and 8), which were flashed on the average at an identical frequency $p = 1/k$, but such that the frequencies of flashes of each light substantively depended on the light flashed immediately preceding it (experiment α). In Fig. 12, as previously, the average reaction time T is shown on the ordinate (obtained from a series of tests, carried out after a prolonged preliminary training of the subjects under controlled conditions in which the individual lights were flashed) and the mean *conditional* entropy on the abscissa :

$$\begin{aligned} H_{\alpha}(\beta) &= p(A_1) H_{A_1}(\beta) + p(A_2) H_{A_2}(\beta) + \dots + p(A_k) H_{A_k}(\beta) \\ &= \frac{1}{k} [H_{A_1}(\beta) + H_{A_2}(\beta) + \dots + H_{A_k}(\beta)] \end{aligned}$$

(A_1, A_2, \dots, A_k being the outcomes of the experiment α). The circumstance that, in Fig. 12, the triangles are found to fall closely along the same straight line, around which the squares and circles are grouped, shows that the conditional entropy $H_{\alpha}(\beta)$ is actually that measure of the amount of uncertainty which determines the dependence of the mean reaction time of the person on the conditions of the experiment.

2.3. The concept of information

We recall the quantity $H(\beta)$ characterizing the amount of uncertainty of an experiment β . When this quantity is 0, it signifies that the outcome of β is known beforehand; the value of $H(\beta)$ being large or small implies that the problem of predicting the result of an experiment is complicated or straightforward, respectively. Some measurement or observation α , preceding an experiment β , may narrow down the number of possible outcomes of β and thereby reduce the amount of its uncertainty; thus, the amount of uncertainty of an experiment, consisting of determining the heaviest of three loads, is reduced after two of

them have been compared by weighing. In order that the result of the measurement (observation) α may yield information about the succeeding experiment β , it is obviously necessary that this result be not known previously; hence, α can be considered as an auxiliary experiment, also having several admissible outcomes. The fact that the realization of α cannot increase the amount of uncertainty of β finds itself reflected in the observation that the conditional entropy $H_\alpha(\beta)$ of β given the occurrence of α is found to be less (more precisely, not greater) than the unconditional entropy $H(\beta)$ of the same experiment. In addition, if the experiment β does not depend on α , then the realization of α does not lower the entropy of β , i.e., $H_\alpha(\beta) = H(\beta)$; if, however, the result of α completely predetermines the outcome of β , then the entropy of β reduces to zero : $H_\alpha(\beta) = 0$. Thus, the difference

$$I(\alpha, \beta) = H(\beta) - H_\alpha(\beta)$$

indicates to what extent the realization of α lowers the uncertainty of β , i.e., how much more we know about the outcome of β by carrying out a measurement (observation) α ; this difference is called the *amount of information with respect to the experiment β , contained in the experiment α* or, briefly, *the information about β contained in α* .

We have thus a *numerical measure of information*, which is extremely fruitful in many cases. Thus, for example, in the conditions of Problem 18 (pp. 66-67) it can be stated that the reaction used yields *information* about the incidence of the subject disease, close to $0.14 - 0.12 = 0.02$ (where we have taken as a unit the information given us by a single 'yes' or 'no' answer to a question, in respect of which we are already inclined to consider an affirmative and negative statement to be equally probable); the digit 0.02 also evaluates the usefulness of the reaction. Other examples of employing the concept of amount of information shall be adduced in Chapters 3 and 4.

The relationship between the concepts of *entropy* and *information* in a well-known sense recalls the relationship between the physical concepts of potential and potential difference. The entropy is an abstract 'measure of uncertainty'; the value of this concept to a considerable extent lies in the fact that it enables us to compute the influence on a specific experiment β of some other experiment α as the 'difference of entropies' $I(\alpha, \beta) = H(\beta) - H_\alpha(\beta)$. Since the concept of information, related to specific changes in the conditions of experiment β , is, so to say, 'more active' than the concept of entropy, hence for imparting a sharper meaning to the entropy it is more expedient to reduce the latter concept to the former one. The entropy $H(\beta)$ of β can be defined as also *the information with respect to β , contained in β itself* (since the realization of the experiment β itself, obviously, completely determines its outcome and, consequently, $H_\beta(\beta) = 0$), or as *the maximum information that can be obtained with respect to β* ('the total information with respect to β '). Differently, the entropy $H(\beta)$ of β

is the information given by the realization of this experiment, i.e., *the average information contained in a single outcome of the experiment* β †. These statements, which will be extensively used in Chapters 3 and 4, have understandably the same meaning as the ‘measure of uncertainty’; the greater the uncertainty of any experiment, the larger is the information obtained by determining its outcome.

We further emphasize that the information, with respect to β , contained in an experiment α is, by definition, the *mean value of the random variable* $H(\beta) - H_{A_i}(\beta)$ associated with the individual outcomes A_i of α ; hence, it can also be termed as ‘the mean information with respect to β contained in α .’ It may often happen that our desire to know the outcome of some experiment β may motivate us to perform an auxiliary experiment (measurement, observation) α which can be selected in a variety of ways; thus, for example, when ascertaining the heaviest of some system of loads, we can compare the individual loads in different orders. In this case, it is recommended to start with that experiment α_0 , which contains the *maximum* information with respect to β , because in a different experiment α it is *likely* that we shall obtain a smaller decrease in the amount of uncertainty of β (the entropy $H(\beta)$). In reality, however, it is also possible that by chance the experiment α occurs to be more useful than α_0 ; in principle, the outcome A of α_0 may turn out to be so unfortunate that the entropy $H_A(\beta)$ is found to be *greater* than the original entropy $H(\beta)$. Such a

†We note that the expression for entropy

$$H(\beta) = -p(B_1) \log p(B_1) - p(B_2) \log p(B_2) - \dots - p(B_l) \log p(B_l)$$

has the form of the mean value of a random variable, taking the values $-\log p(B_1)$, $-\log p(B_2)$, \dots , $-\log p(B_l)$ with probabilities $p(B_1)$, $p(B_2)$, \dots , $p(B_l)$, respectively (see p. 6). In this connection, it may be considered that when a definite outcome B_i of our experiment is realized, we obtain information equal to $-\log p(B_i)$. In such a case, if the experiment β has, say, altogether two possible outcomes B_1 and B_2 with probabilities 0.99 and 0.01, then in realizing the outcome B_1 we obtain quite a small amount of information $-\log 0.99 = 0.017$ bits. This is completely natural; in fact, even prior to the experiment we had known that the outcome B_1 was almost sure to occur, so that the result of experiment makes little change in the information available to us. On the contrary, if the outcome B_2 is realized, then the information obtained equals $-\log 0.01 = 6.6$ bits, i.e., it is much larger than in the first case. This is natural, since the information obtained as a result of the experiment is here of much greater interest (it is the realization of a remotely expected event). However, we seldom obtain such a large amount of information with a large number of repetitions of an experiment. Hence, the *average amount of information* contained in a single outcome of an experiment is found here to be smaller than in the case in which the probability of both outcomes is equal. We further remark that in practical problems we are always interested only in this average amount of information; the idea of the amount of information, related to the individual outcomes of an experiment, is rarely applied.

situation is completely natural, since the random character of the outcomes of β does not obviously permit us to outline in advance the results of this experiment via some shortest route; at most, we can work out and indicate the path, which is found to be *probably* the shortest; it is precisely this possibility which is offered by information theory.[†] The individual quantities $H(\beta) - H_{A_i}(\beta)$ do not factually constitute even the characteristics of the experiment β , because if the result A_i of an experiment α is known to us (and α and β are *not independent*), then we lose the right to speak of the initial experiment β and have to take into account those changes in the conditions of this experiment which stem from the fact that α has an outcome A_i . Thus, $H_{A_i}(\beta)$ is simply the entropy of some *new* experiment to which the experiment β reduces given that the event A_i is realized.

Problem 20. *Suppose that an experiment β consists of the draw of one ball from an urn, containing 5 black and 10 white balls and an experiment α_k consists of the preceding draw of k balls from the same urn (without replacement). Find the entropy of experiment β and the information about this experiment contained in the experiments α_1 , α_2 , α_{13} , and α_{14} ?*

The entropy of β is obviously given by

$$H(\beta) = -\frac{1}{3} \log \frac{1}{3} - \frac{2}{3} \log \frac{2}{3} \approx 0.92 \text{ bits.}$$

Furthermore, by the formulas obtained in the solution of Problem 19, we have (in bits):

$$\begin{aligned} I(\alpha_1, \beta) = H(\beta) - H_{\alpha_1}(\beta) &= -\frac{1}{3} \log \frac{1}{3} - \frac{2}{3} \log \frac{2}{3} \\ &+ \frac{1}{3} \left(\frac{2}{7} \log \frac{2}{7} + \frac{5}{7} \log \frac{5}{7} \right) \\ &+ \frac{2}{3} \left(\frac{5}{14} \log \frac{5}{14} + \frac{9}{14} \log \frac{9}{14} \right) \approx 0.004; \end{aligned}$$

[†]We should not form an impression that the methods of information theory do not always enable us to obtain an absolute evaluation, say, for a number of auxiliary experiments α , needed for determining the result of a definite experiment β . (By absolute evaluation we understand here the evaluation which is not only most probable but has an *absolute* character.) Thus, for instance, if the information $I(\alpha, \beta)$ equals the entropy $H(\beta)$ of experiment β , then we can be convinced that *with every outcome of the experiment α the result of β stands completely defined*. In analogy to this, if the information $I(\alpha, \beta)$ is 0, then with every outcome A_i of the experiment α the entropy $H_{A_i}(\beta)$ equals the original entropy $H(\beta)$. In this connection, see Chapter 3.

$$\begin{aligned}
I(\alpha_2, \beta) = H(\beta) - H_{\alpha_2}(\beta) &= -\frac{1}{3} \log \frac{1}{3} - \frac{2}{3} \log \frac{2}{3} \\
&+ \frac{\binom{10}{2}}{\binom{15}{2}} \left(\frac{5}{13} \log \frac{5}{13} + \frac{8}{13} \log \frac{8}{13} \right) \\
&+ \frac{\binom{10}{1} \binom{5}{1}}{\binom{15}{2}} \left(\frac{4}{13} \log \frac{4}{13} + \frac{9}{13} \log \frac{9}{13} \right) \\
&+ \frac{\binom{5}{2}}{\binom{15}{2}} \left(\frac{3}{13} \log \frac{3}{13} + \frac{10}{13} \log \frac{10}{13} \right) \approx 0.008;
\end{aligned}$$

$$I(\alpha_{13}, \beta) = H(\beta) - H_{\alpha_{13}}(\beta) = -\frac{1}{3} \log \frac{1}{3} - \frac{2}{3} \log \frac{2}{3} - \frac{2 \times 5 \times 10}{15 \times 14} \approx 0.44;$$

and, finally,

$$I(\alpha_{14}, \beta) = H(\beta) - H_{\alpha_{14}}(\beta) = H(\beta) (\approx 0.92).$$

Problem 21. Suppose that the probability that there will or will not be rain at a certain place on 15 June is 0.4 and 0.6, respectively, and on 15 October it is 0.8 and 0.2, respectively. We assume that, following a specified method, the weather forecast on 15 June is found to be correct in $\frac{3}{5}$ (resp. $\frac{4}{5}$) of those cases in which rain (resp. no precipitation) is predicted; when applied to the weather on 15 October this method is found to be correct in $\frac{9}{16}$ (resp. $\frac{1}{2}$) of those cases in which rain (resp. no rain) is predicted (the comparatively higher percentage of error in the latter case is naturally explained by the fact that a low probability event, which is more difficult to guess, is predicted). The question is : On which of the two dates indicated, does the forecast yield us greater information about the actual weather?

We denote by β_1 and β_2 the experiments consisting of the determination of weather at the place under consideration on 15 June and 15 October. We assume that each of these experiments has, in all, two outcomes, B (rain) and \bar{B} (no rain); the corresponding probability tables have the form :

(i) Experiment β_1 :

Outcomes	B	\bar{B}
Probabilities	0.4	0.6

(ii) Experiment β_2 :

Outcomes	B	\bar{B}
Probabilities	0.8	0.2

Consequently, the entropy of experiments β_1 and β_2 is given by

$$H(\beta_1) = -0.4 \log 0.4 - 0.6 \log 0.6 \approx 0.97 \text{ bits},$$

$$H(\beta_2) = -0.8 \log 0.8 - 0.2 \log 0.2 \approx 0.72 \text{ bits}.$$

Now, let α_1 and α_2 be the forecasts of weather on 15 June and 15 October. The experiments α_1 and α_2 also have each two outcomes : A (forecast of rain), \bar{A} (forecast of dry weather); in addition, the pairs of experiments (α_1, β_1) and (α_2, β_2) , are characterized by the accompanying conditional probability tables :

(i) <i>Pair</i> (α_1, β_1) :	$p_A^{(1)}(B)$	$p_{\bar{A}}^{(1)}(\bar{B})$	$p_{\bar{A}}^{(1)}(B)$	$p_A^{(1)}(\bar{B})$
	0.6	0.4	0.2	0.8
(ii) <i>Pair</i> (α_2, β_2) :	$p_A^{(2)}(B)$	$p_{\bar{A}}^{(2)}(\bar{B})$	$p_{\bar{A}}^{(2)}(B)$	$p_A^{(2)}(\bar{B})$
	0.9	0.1	0.5	0.5

(we recall that $p_A(B) + p_A(\bar{B}) = p_{\bar{A}}(B) + p_{\bar{A}}(\bar{B}) = 1$). These tables enable us to determine also the unknown probabilities $p_1(A)$ and $p_1(\bar{A})$; $p_2(A)$ and $p_2(\bar{A})$ of the outcomes A and \bar{A} of experiments α_1 and α_2 . In fact, by the equation of total probability (see p. 23), we have for the experiment β_1

$$0.4 = p(B) = p_1(A) p_A^{(1)}(B) + p_1(\bar{A}) p_{\bar{A}}^{(1)}(B) = 0.6 \times p_1(A) + 0.2 \times p_1(\bar{A}),$$

and for the experiment β_2

$$0.8 = p(B) = p_2(A) p_A^{(2)}(B) + p_2(\bar{A}) p_{\bar{A}}^{(2)}(B) = 0.9 \times p_2(A) + 0.5 \times p_2(\bar{A}).$$

Since $p_1(\bar{A}) = 1 - p_1(A)$, $p_2(\bar{A}) = 1 - p_2(A)$, we obtain

$$p_1(A) = p_1(\bar{A}) = 0.5, \quad p_2(A) = 0.75, \quad p_2(\bar{A}) = 0.25.$$

We now calculate the entropies $H_A(\beta_1)$, $H_{\bar{A}}(\beta_1)$, $H_A(\beta_2)$ and $H_{\bar{A}}(\beta_2)$ (in bits) :

$$H_A(\beta_1) = -0.6 \times \log 0.6 - 0.4 \times \log 0.4 \approx 0.97,$$

$$H_{\bar{A}}(\beta_1) = -0.2 \times \log 0.2 - 0.8 \times \log 0.8 \approx 0.72;$$

and

$$H_A(\beta_2) = -0.9 \times \log 0.9 - 0.1 \times \log 0.1 \approx 0.47,$$

$$H_{\bar{A}}(\beta_2) = -0.5 \times \log 0.5 - 0.5 \times \log 0.5 = 1,$$

Consequently,

$$H_{\alpha_1}(\beta_1) = p_1(A) H_A(\beta_1) + p_1(\bar{A}) H_{\bar{A}}(\beta_1) \approx 0.84.$$

$$H_{\alpha_2}(\beta_2) = p_2(A) H_A(\beta_2) + p_2(\bar{A}) H_{\bar{A}}(\beta_2) \approx 0.60.$$

It is thus seen that the information, contained in the weather forecast for 15 June (experiment α_1) concerning the actual weather on this date (concerning experiment β_1), is given by

$$I(\alpha_1, \beta_1) = H(\beta_1) - H_{\alpha_1}(\beta_1) \approx 0.97 - 0.84 = 0.13 \text{ bits.}$$

This is slightly *greater* than the information concerning the actual weather on 15 October (concerning experiment β_2) contained in the forecast of weather on this date (in experiment α_2) :

$$I(\alpha_2, \beta_2) = H(\beta_2) - H_{\alpha_2}(\beta_2) \approx 0.72 - 0.60 = 0.12 \text{ bits.}$$

This result enables us to consider the forecast of weather on 15 June, to be of greater value than the one on 15 October despite the fact that the latter forecast *more frequently turns out to be correct*; really, by the equation of total probability, the probability that the weather forecast for 15 June would be found correct is

$$p_1(A) p_A^{(1)}(B) + p_1(\bar{A}) p_{\bar{A}}^{(1)}(\bar{B}) = 0.5 \times 0.6 + 0.5 \times 0.8 = 0.7,$$

whereas, for the weather forecast for 15 October, this probability is given by

$$p_2(A) p_A^{(2)}(B) + p_2(\bar{A}) p_{\bar{A}}^{(2)}(\bar{B}) = 0.75 \times 0.9 + 0.25 \times 0.5 = 0.8.$$

In general, the amount of information $I(\alpha, \beta)$, contained in the forecast α about the outcome of some experiment (or random event) β , is the objective characteristic of the value of forecast. It is zero, if $H_{\alpha}(\beta) = H(\beta)$, i.e., if α and β are independent events (so that the 'forecast' α is, in no way, associated with the event β), or if $H(\beta) = 0$ (so that the outcome of β is pre-known and need not be forecast); in all the remaining cases, the amount of information is positive, but is not greater than the amount of uncertainty $H(\beta)$ of the experiment β . (Moreover, $I(\alpha, \beta) = H(\beta)$, only if $H_{\alpha}(\beta) = 0$, i.e., if the 'forecast' α uniquely determines the outcome of β .) However, we note that the universality of the considered method of evaluating the quality of *any* forecast implies that this method cannot cover all aspects of a question. In particular, our estimate of the forecast completely disregards the *contents (meanings)* of various outcomes of the subject experiment β and rests only on *the probability* of these outcomes. However, it seems quite possible in practical life that, owing to the distinct

characters of different outcomes of β , one of them is more crucially important to correct prediction than the others. Thus, when forecasting any natural calamity B (an earthquake, a flood, or even a less hazardous variant, a frost), it is usually of utmost importance that no error be committed in predicting that B *will not occur*, whereas the error in forecasting the *occurrence* of B is most often considerably less grave (it implies only taking unfounded precautionary measures). Such differences among the outcomes of an experiment β have to be taken care of by other numerical characteristics, different from information I .

In this connection, we may reiterate with respect to information I what we stated above (see pp. 54-55) with regard to entropy H . The concept of information first arose directly from the needs of communication theory and is especially oriented to meet the demands of this theory. Since the transmission of message of a specified length over a communication channel (for example, in telegraphy) entails roughly the same amount of time and cost both in the case of completely trivial or even false information as well as in the case of information about a scientific discovery of far-reaching importance, from the viewpoint of communication theory we have to consider that the amount of information in these messages is also identical. Obviously, such a definition of the amount of information, which completely disregards the meaning of its content, may not be appropriate in all the cases in which the term 'information' is employed in our everyday life. It is, however, plain that the value of any scientific concept is determined not by the number of cases it is unable to serve, but mainly by the importance and diffusion of concrete problems, in the solution of which it is found to be fruitful. In relation to the concept of information, such problems are numerous (see, in particular, Chapters 3 and 4).

Problem 22. Suppose that an experiment β consists of determining the position of some point M , relating to which we know beforehand just that it lies on a segment AB of length L (Fig. 13). Let us also suppose that an experiment α consists

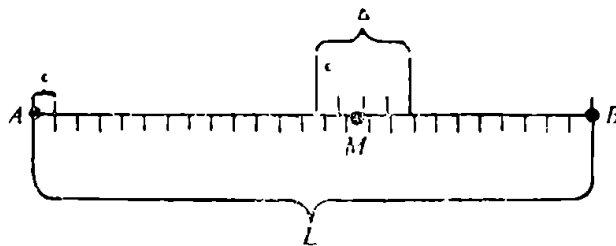


Fig. 13.

of measuring the length of a segment AM by means of some measuring instrument, which gives us the value of length to within a definite 'measurement error' Δ (say, by means of a scale marked with divisions of length Δ). What information about the true position of point M is contained in the result of measurement.

A cursory inspection reveals that this problem cannot be solved with the aid of equations derived above. This is so because at the root of these equations there has always been an experiment that can have only a finite number of outcomes, but here the experiment β has *infinitely many* outcomes (the point M can coincide with any point of the segment AB). And, indeed, we cannot assign here any finite entropy to the experiment β . Nevertheless, it is found that the information $I(\alpha, \beta)$ (which is the difference of two entropies $H(\beta)$ and $H_\alpha(\beta)$) has, in the case under consideration, a completely defined finite value. In order to clarify this, we first assume that the lengths L and Δ are commensurable with each other and split the entire segment AB into small segments of length ϵ , so chosen that an integral number of such small segments lies both on the entire segment AB and a segment of length Δ (i.e., to be such that both the ratios L/ϵ and Δ/ϵ can be expressed by an integer). We now tackle the problem of determining the position of the point M to within the value ϵ . Since we know, beforehand, only that M is placed somewhere on the segment AB , we can consider that an experiment β_ϵ consisting of 'determining the position of M to within ϵ ', has L/ϵ equally probable outcomes. Hence its entropy is $H(\beta_\epsilon) = \log(L/\epsilon)$. Moreover, after carrying out an experiment α , i.e., measuring the length AM , using our measuring instrument, it becomes clear to us that the point M actually lies inside a small interval of length Δ , which determines the accuracy of the measurement. Hence, when the outcome α of experiment β_ϵ is known, we have in all Δ/ϵ equally probable outcomes and, therefore, $H_\alpha(\beta_\epsilon) = \log(\Delta/\epsilon)$. Consequently,

$$I(\alpha, \beta_\epsilon) = H(\beta_\epsilon) - H_\alpha(\beta_\epsilon) = \log \frac{L}{\epsilon} - \log \frac{\Delta}{\epsilon} = \log \frac{L}{\Delta}.$$

For indefinitely decreasing ϵ (i.e., for determining the position of our point with increasing accuracy), both the entropies $H(\beta_\epsilon)$ and $H_\alpha(\beta_\epsilon)$ increase infinitely; however, the information $I(\alpha, \beta_\epsilon)$ is here invariant, always remaining equal to $\log(L/\Delta)$. It is, therefore, natural that the information $I(\alpha, \beta)$ (which we define, say, as the *limit* $I(\alpha, \beta_\epsilon)$ as $\epsilon \rightarrow 0$) should also be considered to be equal to $\log(L/\Delta)$. This number gives us the information relating to the true position of M , contained in the result of measurement to within Δ . For indefinitely increasing accuracy of the instrument (i.e., for indefinitely decreasing Δ), this information increases infinitely, though this increase is comparatively slow: for an n -times increase in accuracy, we obtain in addition only $\log n$ units of information (for example, when the accuracy increases twice, we gain 1 bit of information and when it increases 1000 times, the information gained is less than 10 bits).

In our reasoning, the lengths L and Δ are assumed to be commensurable. It is, however, obvious that this assumption is not essential: if ϵ is chosen sufficiently small, then the assumption that an integral number of small segments of length ϵ are packed on the segments AB and Δ is always satisfied to a great

accuracy, so that the conclusion obtained by us may remain invariant even in case L and Δ are incommensurable.

Here we have simply touched upon the problem of information contained in the result of a measurement. For a more detailed discussion, the reader is referred to Brillouin [5].

We further remark that in solving Problem 22, we encountered a rather unusual situation. We had to deal there with an experiment β having an *infinite number of outcomes*, so that we have to consider the corresponding entropy $H(\beta)$ to be infinite. For computing the information of this experiment, contained in another experiment α , we considered an auxiliary experiment β_ϵ , obtained by combining together the whole group of outcomes of β , differing from each other by a value not larger than a small ϵ . It was also found that both the entropy $H(\beta_\epsilon)$ of this new experiment and the conditional entropy $H_\alpha(\beta_\epsilon)$ have a finite value; furthermore, since their difference was found to be independent of the choice of ϵ , we agreed to take this difference also as the information $I(\alpha, \beta)$.

A similar sort of situation continues to recur whenever we consider an experiment β having a continuous set of outcomes. In all such cases, the entropy $H(\beta)$ is infinite. However, in place of it we may often consider a finite entropy $H(\beta_\epsilon) = H_\epsilon(\beta)$, obtained by combining together all outcomes of β , differing not more than by some small ϵ . In practical problems, the entropy $H_\epsilon(\beta)$ (called the *ϵ -entropy of an experiment β*) represents a quite reasonable quantity, since we cannot, in general, distinguish the outcomes of β that differ from each other by less than a definite very small value. This value is determined by the limiting accuracy of a measuring instrument at our disposal. We shall take up this problem again later (see Chapter 4.3).

Equating the entropy $H(\alpha)$ to the average information contained in the outcome of an experiment α , we can, in particular, impart a new interpretation to the psychological experiments described on pp. 56-59 and 72-73. It is now seen that, according to these results, the mean time required for an accurate understanding of the meaning of some signal, and the proper reaction to it, increases in proportion to the mean information contained in this signal. It is natural to assume on this basis that in the case in which the events occur with sufficient regularity, in other words, are characterized by a definite statistical law (i.e., are random events in a strict sense of the probability theory), the information on the emergence of such an event is conveyed by the sensory organs and nervous system over time that is on the average proportional to the amount of information contained in this event. Hence, it can be assumed that the transmission of information in living organisms is characterized in many cases by the following property: *the same amount of information is transmitted on the average over the same period of time*. It is worth noting here that, as we shall see from the contents of Chapter 4, such a property holds also in the transmission of information over all engineering communication lines.

There is a simple consequence from the assumption made and this can be verified experimentally. Let us presume that, while carrying out an experiment to determine the average reaction time, the subject is forced to react quite fast, so fast that he is himself unable to comprehend fully what signal precisely appeared before him. For example, let the signal we consider consist of the flashing of one of n lights and let it be required that the i th knob is pressed when i th light is flashed. If the subject is forced to decrease reaction time T , he naturally errs with greater frequency, pressing in place of the i th knob, some other knob, say the j th one. This means that, because of the compulsion to react very fast, he is not in a position to absorb fully all the information included in the appearance of a specific signal. If, however, T is not too small, then the subject is able to grasp some useful information about the signal. This is manifested by the fact that his reactions are not completely disorderly, rather, on the average, it is oftner the i th knob that he presses than any other one when the i th light is flashed. An experiment α consisting of pressing one of the n knobs by the subject contains here definite information about an experiment β consisting in the flashing of one of the n lights. This information $I(\alpha, \beta)$ is obviously the average information about β that the subject is able to comprehend in time T . According to our assumption, the dependence of this information on the reaction time T must be the same as that of the entropy $H(\beta)$ on T for the case in which T is defined as the least time sufficient for an *error-free* reaction.

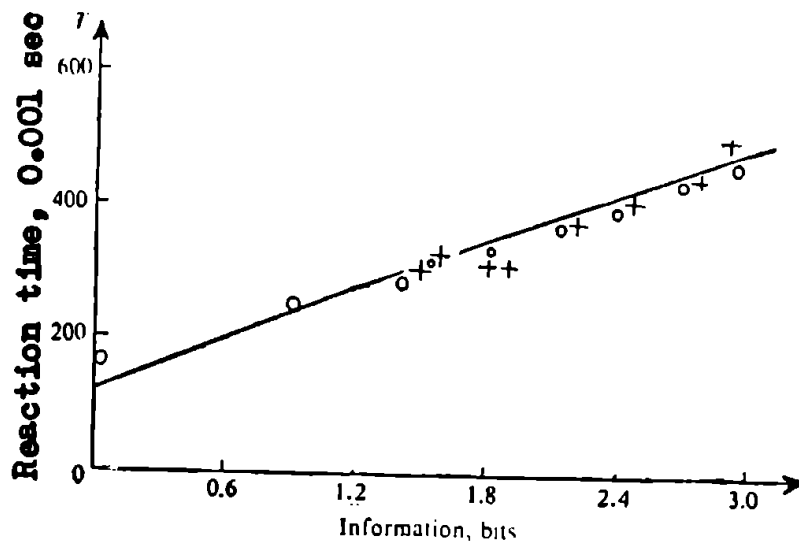


Fig. 14.

The last conclusion has been verified by the British psychologist, W. E. Hick [46]. The results obtained by him are plotted in Figure 14. The small circles denote here the average time determined from the experiments of the same kind as described on pp. 56-58. Specifically, before the subject (who is the investigator himself in the given case) there are flashed, with equal frequency, n distinct lights (n ranging from 1 to 10 in different experiments) and the average time T , requisite for correct reaction to the incoming signal, is measured. As known already,

T increases here linearly with the growth of the entropy $H(\beta) = I(\beta, \beta)$. This is manifested by the fact that in Fig. 14, where T is plotted on the ordinate and $H(\beta) = I(\beta, \beta)$ on the abscissa, all circles fall quite accurately along a single straight line. The crosses in Fig. 14 plot the results of an experiment in which all 10 lights are used and flashed at the same frequency, but the reaction time T is fixed beforehand to be so small that the reaction of the subject in a series of cases is necessarily found to be faulty. In order to evaluate the average information contained in an experiment α (the pressing of one of the 10 knobs by the subject) about an experiment β (the emergence of one of the 10 signals), a large series of experiments N was carried out with one and the same T and $n_{i,j}$ was calculated, being the total number of all those cases in which the j th knob was pressed in response to the flash of the i th lamp. Here i and j take all possible values from 1 to 10 and the sum of all $n_{i,j}$ is equal to N , where N is the total number of experiments, while the total number of all cases in which the subject reacted correctly is given by $n_{1,1} + n_{2,2} + \dots + n_{10,10}$. It is clear that the probability of 10 outcomes of experiment β may be considered here to be given approximately by the frequencies

$$q_1 = \frac{n_{1,1} + n_{1,2} + \dots + n_{1,10}}{N},$$

$$q_2 = \frac{n_{2,1} + n_{2,2} + \dots + n_{2,10}}{N}, \dots, q_{10} = \frac{n_{10,1} + n_{10,2} + \dots + n_{10,10}}{N},$$

and the probability of 10 outcomes of experiment α by the frequencies

$$p_1 = \frac{n_{1,1} + n_{2,1} + \dots + n_{10,1}}{N}$$

$$p_2 = \frac{n_{1,2} + n_{2,2} + \dots + n_{10,2}}{N}, \dots, p_{10} = \frac{n_{1,10} + n_{2,10} + \dots + n_{10,10}}{N}.$$

The compound experiment $\alpha\beta$ has here $10^2 = 100$ different outcomes, whose probabilities are approximately equal to the following frequencies

$$p_{1,1} = \frac{n_{1,1}}{N}, p_{1,2} = \frac{n_{1,2}}{N}, \dots, p_{10,10} = \frac{n_{10,10}}{N}.$$

This yields the following expressions for the entropies of experiments β , α and $\alpha\beta$:

$$H(\beta) = -q_1 \log q_1 - q_2 \log q_2 - \dots - q_{10} \log q_{10},$$

$$H(\alpha) = -p_1 \log p_1 - p_2 \log p_2 - \dots - p_{10} \log p_{10},$$

$$H(\alpha\beta) = -p_{1,1} \log p_{1,1} - p_{1,2} \log p_{1,2} - \dots - p_{10,10} \log p_{10,10},$$

which permit us to calculate approximately these entropies by the experimentally

determined numbers $n_{i,j}$. Then, from the formula

$$H(\alpha\beta) = H(\alpha) + H_{\alpha}(\beta)$$

(see p. 63) the mean conditional entropy $H_{\alpha}(\beta)$ can be defined as

$$H_{\alpha}(\beta) = H(\alpha\beta) - H(\alpha).$$

Moreover, by $H(\beta)$ and $H_{\alpha}(\beta)$ we can also determine the information $I(\alpha, \beta)$ about the experiment β , contained in the experiment α :

$$I(\alpha, \beta) = H(\beta) - H_{\alpha}(\beta).$$

This value of $I(\alpha, \beta)$ is used as the abscissa of crosses in Fig. 14.

We note that the setting of the experiment here is in a certain sense converse to that considered on pp. 56-59 and 72-73. Earlier, the information $I(\beta, \alpha) = H(\beta)$ was defined beforehand and we investigated the dependence of reaction time T on it. Against this, T is now specified beforehand (i.e., it is required for the subject to react over a definite time T after the emergence of signals) and the dependence of information $I(\alpha, \beta)$ on this is studied. The circumstance that the crosses in Figure 14 cluster around the same straight line as the circles do affirms the surmise that the reaction time is linearly dependent precisely on the *information* contained in a signal.

There is obviously no justification to extend the results of these few experiments, carried out in a highly specific set up, to all general processes of the transmission of information in a living organism. In fact, a simple linear dependence between the reaction time and the information contained in a given signal is not observed in all experiments. Besides, even in those cases in which such a dependence holds, the coefficients of corresponding linear functions may take highly different values (see, for example, Fig. 15 taken from Nikolaev [51]; also, see [52] and the book [50] containing more than 500 references). The factors on which these coefficients depend have been studied by many authors (see, for example, the review papers [48] and [49]); in this field, there still remain a large number of open questions, however. Nonetheless, the available data (see references cited above and also [42] and [19]) show positively that the quantitative concept of information can often be used successfully to give a mathematical description of the processes of perception and assimilation of various sorts of signals by living organisms that are transmitted to the organisms from the external world.

We shall now show that *the information with respect to an experiment β contained in an experiment α is always the same as the information with respect to α contained in β* . This is immediate from the equations of preceding section: since

$$H(\alpha) + H_{\alpha}(\beta) = H(\beta) + H_{\beta}(\alpha)$$

(see p. 66), it follows that

$$I(\alpha, \beta) = H(\beta) - H_{\alpha}(\beta) = H(\alpha) - H_{\beta}(\alpha) = I(\beta, \alpha).$$

Thus, the information $I(\alpha, \beta)$ that α contains with respect to β can also be called *the reciprocal information of two experiments α and β with respect to each other.*

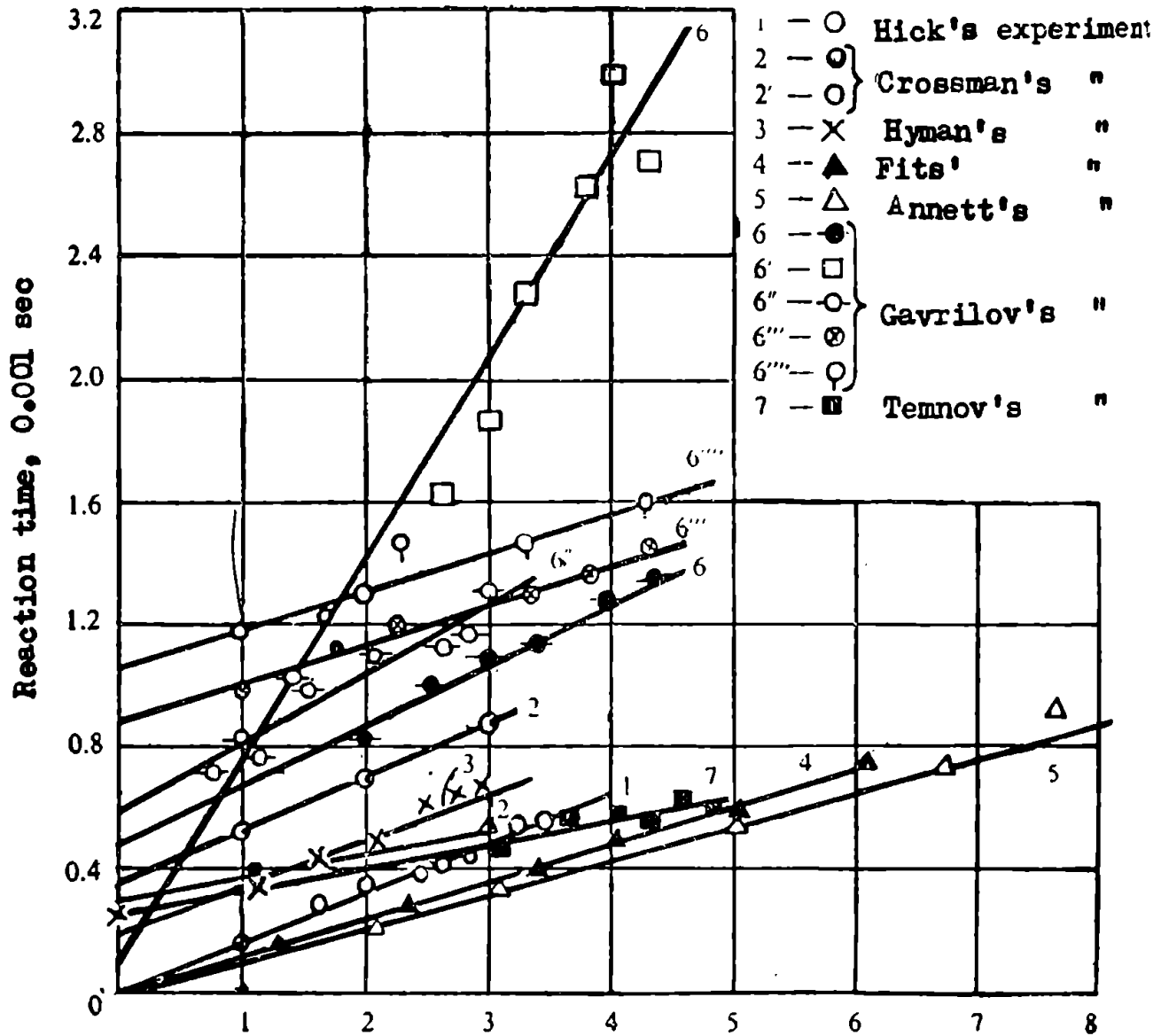


Fig. 15.

The equality between information $I(\alpha, \beta)$ and information $I(\beta, \alpha)$ is emphasized by the following simple equation, which is found to be extremely convenient in many cases:

$$I(\alpha, \beta) = H(\alpha) + H(\beta) - H(\alpha\beta)$$

(see, for example, p. 85). This equation stems from the fact that $H_{\alpha}(\beta) = H(\alpha\beta) - H(\alpha)$ (because $H(\alpha\beta) = H(\alpha) + H_{\alpha}(\beta)$); the experiments α and β entering the right-hand side of this equation are completely symmetric.

The symmetric equation derived here for the amount of information can also be usefully transformed. This transformation simplifies its right-hand side, so that it can be expressed directly in terms of the probabilities $p(A_1), \dots, p(A_k), p(B_1), \dots, p(B_l)$ and $p(A_1B_1), p(A_1B_2), \dots, p(A_kB_l)$ of distinct outcomes of α, β and $\alpha\beta$. In fact, according to the definition of entropy,

$$H(\alpha) = -p(A_1) \log p(A_1) - p(A_2) \log p(A_2) - \dots - p(A_k) \log p(A_k),$$

$$H(\beta) = -p(B_1) \log p(B_1) - p(B_2) \log p(B_2) - \dots - p(B_l) \log p(B_l).$$

and

$$H(\alpha\beta) = -p(A_1B_1) \log p(A_1B_1) - p(A_1B_2) \log p(A_1B_2) - \dots \\ - p(A_kB_l) \log p(A_kB_l).$$

On the other hand, by the addition law of probabilities (see p. 9),

$$p(A_i) = p(A_iB_1) + p(A_iB_2) + \dots + p(A_iB_l), \quad i = 1, 2, \dots, k,$$

and

$$p(B_j) = p(A_1B_j) + p(A_2B_j) + \dots + p(A_kB_j), \quad j = 1, 2, \dots, l,$$

so that

$$\begin{aligned} -p(A_i) \log p(A_i) &= -p(A_iB_1) \log p(A_i) - p(A_iB_2) \log p(A_i) - \dots \\ &\quad - p(A_iB_l) \log p(A_i), \\ -p(B_j) \log p(B_j) &= -p(A_1B_j) \log p(B_j) - p(A_2B_j) \log p(B_j) - \dots \\ &\quad - p(A_kB_j) \log p(B_j). \end{aligned}$$

Substituting all these expressions in the original equation, we get

$$\begin{aligned} I(\alpha, \beta) &= -p(A_1B_1) [\log p(A_1) + \log p(B_1) - \log p(A_1B_1)] \\ &\quad - p(A_1B_2) [\log p(A_1) + \log p(B_2) - \log p(A_1B_2)] \\ &\quad \dots \dots \dots \\ &\quad - p(A_kB_l) [\log p(A_k) + \log p(B_l) - \log p(A_kB_l)], \end{aligned}$$

or, finally,

$$\begin{aligned} I(\alpha, \beta) &= p(A_1B_1) \log \frac{p(A_1B_1)}{p(A_1)p(B_1)} + p(A_1B_2) \log \frac{p(A_1B_2)}{p(A_1)p(B_2)} + \dots \\ &\quad + p(A_kB_l) \log \frac{p(A_kB_l)}{p(A_k)p(B_l)}. \end{aligned}$$

This equation also is obviously symmetric in α and β .

The equation

$$I(\alpha, \beta) = I(\beta, \alpha)$$

can also be written in the following form:

$$I(\alpha, \beta) = H(\alpha) - H_{\beta}(\alpha).$$

From this it follows that the *information* $I(\alpha, \beta)$ contained in an experiment α with respect to an experiment β does not exceed the entropy $H(\alpha)$ of α , a fact that is often found useful. This premise obviously cannot be considered as something unexpected. It is natural that the information that α contains *about another experiment* β does not exceed the information contained in α with respect to *itself*, the entropy $H(\alpha)$ of this experiment. Thus, the entropy $H(\alpha)$ can also be defined as the *maximum information which can be contained in an experiment* α (the 'total information' contained in α).

From the formula $I(\alpha, \beta) = H(\alpha) - H_{\beta}(\alpha)$ it also follows that *the information* $I(\alpha, \beta)$ *is precisely equal to the entropy* $H(\alpha)$ *of* α *if and only if the conditional entropy* $H_{\beta}(\alpha)$ *is 0, i.e., if the result of experiment* β *completely determines the outcome of the auxiliary experiment* α . The position will be precisely so, for instance, in the problems analyzed in the next chapter. If, however, $H_{\beta}(\alpha) \neq 0$, then *the information* $I(\alpha, \beta)$ *equals the entropy* $H(\alpha)$ *minus the value* $H_{\beta}(\alpha)$. In particular, *if the experiments* α *and* β *are independent (and only in that case),* $I(\alpha, \beta)$ *is 0.*

We further note that, if the conditional entropy $H_{\beta}(\alpha)$ is 0 and, consequently, the information $I(\alpha, \beta)$ with respect to β , contained in α , is the maximum (i.e., the experiment α does not contain more information about any other experiment β_1), then *the information with respect to every experiment* γ *independent of* β , *contained in* α , *is 0.* This provides the justification to say that the experiment α is 'directed straight' at elucidating the outcome of β and does not contain any 'extraneous' information. In the general case, however, *the information with respect to any experiment* γ *independent of* β , *contained in* α , *does not exceed the quantity* $H_{\beta}(\alpha) = I(\alpha, \alpha) - I(\alpha, \beta)$ (if $H_{\beta}(\alpha) = 0$, then this statement converts into the more particular result indicated above). The proof of the statement made demands the introduction of an important auxiliary concept; it will be adduced (together with the proof of other statement formulated below) at the end of the present section.

We now suppose that α , β and γ are three arbitrary experiments. In such a case, we always have

$$I(\beta\gamma, \alpha) \geq I(\beta, \alpha);$$

in other words, *the information contained in a compound experiment* $\beta\gamma$ (i.e., a pair of experiments β and γ) *with respect to every experiment* α *is never less than that contained in a simple experiment* β . This fact is completely natural from the viewpoint of our heuristic notions on 'information'; a rigorous proof of this and similar propositions provides a justification for the use of the term 'information' in relation to the quantity $I(\alpha, \beta)$. In addition, the equality $I(\beta\gamma, \alpha) = I(\beta, \alpha)$ holds if and only if the conditional probability of any outcome of α , given

that β and γ have certain definite outcomes, remains invariant for a change in the outcome of γ (i.e., it depends only on the outcome of β). In the latter case, it is quite natural to consider that the compound experiment $\beta\gamma$ contains no additional information with respect to α in comparison with the experiment β , so that the equality $I(\beta\gamma, \alpha) = I(\beta, \alpha)$ is also in full agreement here with the intuitive meaning of the concept of 'information.'

Let us now assume that the equality $I(\beta\gamma, \alpha) = I(\beta, \alpha)$ holds. It can be shown that, in this case, we always have

$$I(\gamma, \alpha) \leq I(\beta, \alpha).$$

Thus, if the compound experiment $\beta\gamma$ contains no additional information about α in comparison with the experiment β , then the information about α contained in γ cannot be greater than that contained in β . In addition, the 'less than or equal to' sign in the last inequality can be replaced by the 'equality' sign if and only if $I(\beta\gamma, \alpha) = I(\gamma, \alpha)$, i.e., if the compound experiment $\beta\gamma$ does not contain additional information about α also in comparison with the experiment γ .

The inequality $I(\gamma, \alpha) \leq I(\beta, \alpha)$, referred to above, plays a significant role in information theory (see, for example, [14] and [44] as well as Chapter 4 of this book). It says that in *successive transmissions of information* about an experiment α realized by a chain of experiments $\beta, \gamma, \delta, \dots$, where only β is directly associated with α but γ receives all of the information contained in β about α from its association with β (so that $\beta\gamma$ contains no additional information about α as compared with β), δ receives all of the information about α from its association with the experiment γ and so on, the information about α alone can only be reduced:

$$H(\alpha) = I(\alpha, \alpha) \geq I(\beta, \alpha) \geq I(\gamma, \alpha) \geq I(\delta, \alpha) \geq \dots$$

As an auditory illustration of this situation, we can consider the well-known children's game of a 'garbled telephone'. In this game, the first player utters quietly into the ear of his immediate neighbour some word (the experiment α), the neighbour quietly conveys the word heard by him (which may also differ from the one pronounced originally) to the next player (the experiment β), this player also conveys the word heard by him to his immediate neighbour (the experiment γ), and so on. At the close of the game, each player tells what word he heard and among the participants the one who was first to hear incorrectly the word conveyed to him is regarded to be the loser. In this game, it may so happen that the second player conveys the originally spoken word incorrectly but the third, in consequence of another error, says that he heard the same word as that conveyed in the beginning; however, when this procedure is repeated a large number of times, the second player certainly conveys the word uttered by the first player *on the average* more often than the third player. But our concept of information I is precisely also a statistical concept, characterizing rela-

tions that hold 'on the average'; hence the whole string of inequalities set forth above are always satisfied. It is clear that, from the viewpoint of the intuitive notions on the transmission of information, this situation can also be considered as obvious.

The inequalities

$$I(\beta\gamma, \alpha) \geq I(\beta, \alpha) \quad \text{and} \quad I(\beta\gamma, \alpha) > I(\gamma, \alpha)$$

(see p. 88) can be augmented by one more inequality that is somewhat less obvious from the viewpoint of the intuitively expected properties of the quantity given the name 'information.' It is clear that, in general, it is completely plausible for the inequality

$$I(\beta\gamma, \alpha) < I(\beta, \alpha) + I(\gamma, \alpha)$$

to hold. In fact, if, say, $\beta = \gamma$, then also $\beta\gamma = \beta$, and hence usually in such a case

$$I(\beta\gamma, \alpha) = I(\beta, \alpha) < I(\beta, \alpha) + I(\gamma, \alpha) = 2I(\beta, \alpha).$$

If, however, the experiments β and γ are *independent* (i.e., $I(\beta, \gamma) = I(\gamma, \beta) = 0$), then the inequality $I(\beta\gamma, \alpha) < I(\beta, \alpha) + I(\gamma, \alpha)$ is impossible; in this case, we necessarily have

$$I(\beta\gamma, \alpha) \geq I(\beta, \alpha) + I(\gamma, \alpha).$$

The inequality $I(\beta, \alpha) + I(\gamma, \alpha) > I(\beta\gamma, \alpha)$ being impossible here is explained by the fact that the independence of experiments β and γ guarantees the absence of a 'common portion' of the quantities $I(\beta, \alpha)$ and $I(\gamma, \alpha)$. In fact, here experiments β and γ supply substantially *different* information about the experiment α and therefore the information $I(\beta\gamma, \alpha)$ associated with the simultaneous realization of both the experiments β and γ cannot be less than the sum of $I(\beta, \alpha)$ and $I(\gamma, \alpha)$. This can be compared with the inequality

$$\text{area}(F_1 + F_2) < \text{area } F_1 + \text{area } F_2,$$

where $F_1 + F_2$ is the *union* of figures F_1 and F_2 . This inequality is obviously impossible if F_1 and F_2 do not have a common part. However, it seems that here we may expect the equality

$$I(\beta\gamma, \alpha) = I(\beta, \alpha) + I(\gamma, \alpha),$$

because it remains obscure as to owing to what circumstances $I(\beta\gamma, \alpha)$ can be found to be *greater than* the sum of $I(\beta, \alpha)$ and $I(\gamma, \alpha)$.

The matter, however, is that even for the case in which β and γ are independent, their joint occurrence (i.e., the experiment $\beta\gamma$), which enables us to know *simultaneously* the outcomes of both β and γ , can generally supply more information than that given by the individual realizations of β and γ (with which the quantity $I(\beta, \alpha) + I(\gamma, \alpha)$ is associated). This can be illustrated by the example printed in small type on pp. 25-26. We recall the tetrahedron of Fig. 2 and suppose that the experiments α , β and γ consist of verifying, respectively, that the digits 1, 2 and 3 are or are not on the same side on which the tetrahedron falls. In this case, α can have the outcomes A and \bar{A} , β the outcomes B and \bar{B} and γ the outcomes C and \bar{C} . From the calculations derived on p. 25 it is immediate that α , β and γ are all independent. Thus, we have

$$I(\beta, \alpha) = 0 \quad \text{and} \quad I(\gamma, \alpha) = 0, \quad \text{so that} \quad I(\beta, \alpha) + I(\gamma, \alpha) = 0.$$

On the other hand, the results of the compound experiment $\beta\gamma$ completely determine the outcome of α . (In fact, experiment α has an outcome A if β and γ have a 'common' outcome, i.e.,

both β and γ have, respectively, the 'positive' outcomes B and C , or even 'negative' outcomes \bar{B} and \bar{C} ; α has an outcome \bar{A} if β and γ have 'different' outcomes, i.e., B and \bar{C} or \bar{B} and C . Thus, we have

$$I(\beta\gamma, \alpha) = H(\alpha) = 1 \text{ bit,}$$

i.e.,

$$I(\beta\gamma, \alpha) > I(\beta, \alpha) + I(\gamma, \alpha) \quad (= 0).$$

Furthermore, here experiments β and γ contain *no* information about α , but experiment $\beta\gamma$ contains 'complete' information about α , i.e., the *maximum* information one can have about α .

The proof of the statements made above can be deduced by studying the quantity

$$I_{\beta}(\gamma, \alpha) = H_{\beta}(\alpha) - H_{\beta\gamma}(\alpha),$$

which we call the *mean conditional information of two experiments γ and α with respect to each other, given that experiment β is realized*, or, for short, simply *the conditional information of experiments γ and α given β* . We first note that the conditional information $I_{\beta}(\gamma, \alpha)$ is always *non-negative*. This fact straightaway stems from the inequality

$$H_{\beta\gamma}(\alpha) \leq H_{\beta}(\alpha),$$

signifying that the prior realization of a compound experiment $\beta\gamma$ (i.e., the two experiments β and γ) always reduces the amount of uncertainty of the experiment α to an extent not less than the realization of a single experiment β (for a rigorous proof of this inequality, see Appendix I at the end). Since, in addition, we always have $H_{\beta\gamma}(\alpha) \geq 0$ (because $H_{\beta\gamma}(\alpha)$ is some conditional entropy), hence

$$0 \leq I_{\beta}(\gamma, \alpha) \leq H_{\beta}(\alpha).$$

Moreover, $I_{\beta}(\gamma, \alpha) = H_{\beta}(\alpha)$ if and only if $H_{\beta\gamma}(\alpha) = 0$, i.e., if the compound experiment $\beta\gamma$ uniquely determines the outcome of experiment α ; $I_{\beta}(\gamma, \alpha) = 0$ if and only if $H_{\beta\gamma}(\alpha) = H_{\beta}(\alpha)$ and, consequently, also $I(\beta\gamma, \alpha) = I(\beta, \alpha)$, i.e., if the conditional probabilities of all outcomes of α , given that β and γ have some specific outcomes, do not depend on the outcome of γ (see the end of Appendix I).

We shall now show that the conditional information has *symmetry* property :

$$I_{\beta}(\gamma, \alpha) = I_{\beta}(\alpha, \gamma)$$

(this property is emphasized by the very name 'conditional information of experiments γ and α with respect to each other'). In fact, by definition

$$I_{\beta}(\gamma, \alpha) = H_{\beta}(\alpha) - H_{\beta\gamma}(\alpha), \quad I_{\beta}(\alpha, \gamma) = H_{\beta}(\gamma) - H_{\alpha\beta}(\gamma).$$

But the compound experiment $\alpha\beta\gamma$, consisting of the realization of three experiments α , β and γ , can be considered with equal justification to be a union of the joint experiment $\alpha\beta$ and experiment γ , or also as the union of α and the joint experiment $\beta\gamma$.† Hence,

$$H(\alpha\beta\gamma) = H(\alpha\beta) + H_{\alpha\beta}(\gamma) = H(\beta) + H_{\beta}(\alpha) + H_{\alpha\beta}(\gamma),$$

†Symbolically, this can be written as the equation

$$\alpha\beta\gamma = (\alpha\beta)\gamma = \alpha(\beta\gamma)$$

(cf. the 'associative law' of multiplication of events on p. 38, Chap. 1.5).

and

$$H(\alpha\beta\gamma) = H(\beta\gamma) + H_{\beta\gamma}(\alpha) = H(\beta) + H_{\beta}(\gamma) + H_{\beta\gamma}(\alpha).$$

Consequently,

$$H_{\beta}(\alpha) + H_{\alpha\beta}(\gamma) = H_{\beta}(\gamma) + H_{\beta\gamma}(\alpha),$$

i.e.,

$$I_{\beta}(\gamma, \alpha) = H_{\beta}(\alpha) - H_{\beta\gamma}(\alpha) = H_{\beta}(\gamma) - H_{\alpha\beta}(\gamma) = I_{\beta}(\alpha, \gamma).$$

The equality $I_{\beta}(\gamma, \alpha) = I_{\beta}(\alpha, \gamma)$ is also implied by the following 'symmetry expression' of conditional information $I_{\beta}(\gamma, \alpha)$, which is trivial to verify directly: If A_i (where $i = 1, 2, \dots, l$), B_j (where $j = 1, 2, \dots, m$) and C_k (where $k = 1, 2, \dots, n$) are all possible outcomes of experiments α, β and γ , then

$$I_{\beta}(\gamma, \alpha) = p(B_1) I_{B_1}(\gamma, \alpha) + p(B_2) I_{B_2}(\gamma, \alpha) + \dots + p(B_m) I_{B_m}(\gamma, \alpha).$$

Here

$$I_{B_j}(\gamma, \alpha) = p_{B_j}(A_1 C_1) \log \frac{p_{B_j}(A_1 C_1)}{p_{B_j}(A_1) p_{B_j}(C_1)} + \dots + p_{B_j}(A_l C_n) \log \frac{p_{B_j}(A_l C_n)}{p_{B_j}(A_l) p_{B_j}(C_n)}$$

is reciprocal information of experiments α and γ , given that experiment β has outcome B_j . Such an expression neatly explains the meaning of the conditional information $I_{\beta}(\gamma, \alpha)$; this shall not be needed by us, however.

From the equation $I_{\beta}(\gamma, \alpha) = H_{\beta}(\alpha) - H_{\beta\gamma}(\alpha)$ it is easy to obtain the important relation

$$I(\beta\gamma, \alpha) = I(\beta, \alpha) + I_{\beta}(\gamma, \alpha),$$

close in form to the equation $H(\beta\gamma) = H(\beta) + H_{\beta}(\gamma)$. (The stated relation for $I(\beta\gamma, \alpha)$ stems from the fact that $I(\beta\gamma, \alpha) = H(\alpha) - H_{\beta\gamma}(\alpha)$ and $I(\beta, \alpha) = H(\alpha) - H_{\beta}(\alpha)$.) It is obvious that our assertions concerning the amount of information $I(\beta\gamma, \alpha)$ are automatic consequences of this relation and the properties of conditional information.

In the sequel, we shall also find fruitful the following *triple information equation*:

$$I(\beta\gamma, \alpha) + I(\beta, \gamma) = I(\alpha\gamma, \beta) + I(\alpha, \gamma).$$

For proof of this equation, it is necessary just to interchange the places of β and γ in the expression obtained for $I(\beta\gamma, \alpha)$ and use the analogous expression for $I(\alpha\gamma, \beta)$. By carrying out this procedure, we obtain the same expression on the right- and left-hand sides of our formula

$$I(\beta\gamma, \alpha) + I(\beta, \gamma) = I(\gamma, \alpha) + I_{\gamma}(\beta, \alpha) + I(\beta, \gamma),$$

and

$$I(\alpha\gamma, \beta) + I(\alpha, \gamma) = I(\gamma, \beta) + I_{\gamma}(\alpha, \beta) + I(\alpha, \gamma).$$

From the triple information equation we obtain directly the conclusion indicated above about the amount of information with respect to γ contained in experiment α , when γ is independent of some other experiment β . In fact, the independence of β and γ implies that $I(\beta, \gamma) = 0$; on the other hand, we know that $I(\alpha\gamma, \beta) \geq I(\alpha, \beta)$ always. By virtue of the triple information equation it therefore follows, with β and γ independent, that

$$I(\alpha, \gamma) = I(\beta\gamma, \alpha) - I(\alpha\gamma, \beta) \leq I(\beta\gamma, \alpha) - I(\alpha, \beta) = I_{\beta}(\gamma, \alpha);$$

moreover, $I_{\beta}(\gamma, \alpha)$ is never greater than $H_{\beta}(\alpha)$. On the other hand, making use of the ‘symmetry’ property of information (i.e., the equality $I(\alpha, \beta) = I(\beta, \alpha)$), we can rewrite the triple information equation as

$$I(\beta\gamma, \alpha) + I(\beta, \gamma) = I(\beta, \alpha\gamma) + I(\gamma, \alpha),$$

and replace the inequality $I(\alpha\gamma, \beta) \geq I(\alpha, \beta)$ by the inequality

$$I(\beta, \alpha\gamma) \geq I(\beta, \alpha).$$

This implies straightaway that, with β and γ independent (i.e., with $I(\beta, \gamma) = 0$),

$$I(\beta\gamma, \alpha) \geq I(\beta, \alpha) + I(\gamma, \alpha)$$

(see p. 90).

The inequality $I(\gamma, \alpha) \leq I(\beta, \alpha)$ with $I_{\beta}(\gamma, \alpha) = 0$ can also be obtained from the triple information equation. The derivation can be made if we only replace $I(\alpha\gamma, \beta)$ in this equation by $I(\gamma, \beta) + I_{\gamma}(\alpha, \beta)$ and note that in this case $I(\beta\gamma, \alpha) = I(\beta, \alpha)$, and that the information always has the symmetry property. Then, we arrive at the relation

$$I(\beta, \alpha) = I(\gamma, \alpha) + I_{\gamma}(\alpha, \beta),$$

proving that our inequality holds. It is also observed that this inequality becomes an equality if and only if $I_{\gamma}(\alpha, \beta) = 0$. In this case $I(\gamma, \alpha) = I(\beta\gamma, \alpha)$, i.e., the compound experiment $\beta\gamma$ contains no additional information with respect to α in comparison with γ , a situation we had noted earlier also.

Finally, we recall the fact that the inequality $I(\beta\gamma, \alpha) \geq I(\beta, \alpha)$ (which says that ‘the information contained in a compound experiment $\beta\gamma$ about any experiment α is not less than that contained in a simple experiment β ’) can be associated in a sense with the entropy inequality $H(\beta\gamma) \geq H(\beta)$ (‘the amount of uncertainty of a joint experiment $\beta\gamma$ is never less than that of a simple experiment β ’). However, in the entropy case, there is also one more estimate of the quantity $H(\beta\gamma)$ in a different direction: $H(\beta\gamma) \leq H(\beta) + H(\gamma)$ (‘the amount of uncertainty of a compound experiment $\beta\gamma$ is never greater than the sum of the uncertainties of the individual experiments β and γ ’). In the case of information, the position is rather different: knowing the amount of information about an experiment α that is contained in two experiments β and γ , we cannot estimate from the above the information concerning α that is contained in a compound experiment $\beta\gamma$. Thus, in the case analyzed on pp. 90-91 (where the experiments α , β and γ consist of determining that the digits 1, 2 and 3, respectively, appear on the side on which the tetrahedron of Fig. 2 falls), we would have

$$I(\beta, \alpha) = I(\gamma, \alpha) = 0, \text{ but } I(\beta\gamma, \alpha) = 1 (= H(\alpha)).$$

Hence, from the fact that $I(\beta, \alpha)$ and $I(\gamma, \alpha)$ are both small, it is impossible to infer that $I(\beta\gamma, \alpha)$ is small, too.

2.4. Entropy (revisited). The determination of entropy from its properties

The central theme of this chapter is the concept of entropy as a measure of the uncertainty of an experiment α having random outcomes. The motivation of Section 2.1 was to explain how the conventional definition of entropy is ‘natural’; however, the corresponding arguments were only of a leading nature. The real justification for such a definition of the measure of uncertainty is provided by the whole string of theorems proved in this chapter and Chapter 4, as well as in Appendix I. We shall now recall the definition of entropy and show that it

necessarily stems from the elementary requirements naturally imposed on a quantity that is called upon to serve as the quantitative measure of the amount of uncertainty.

It is natural to assume that the measure of the amount of uncertainty $H(\alpha)$ (which we call entropy) of an experiment α with the accompanying probability table :

<i>Outcomes of experiment</i>	A_1	A_2	\dots	A_k
<i>Probabilities</i>	$p(A_1)$	$p(A_2)$	\dots	$p(A_k)$

must depend only on the variables $p(A_1), p(A_2), \dots, p(A_k)$ (i. e., it is a function of these variables). We denote here the probabilities $p(A_1), p(A_2), \dots, p(A_k)$ by p_1, p_2, \dots, p_k and the entropy $H(\alpha)$ by $H(p_1, p_2, \dots, p_k)$ (see p. 50).

We now formulate those conditions that are naturally required to be satisfied by the function $H(p_1, p_2, \dots, p_k)$. In the first place, it is plain that this function does not have to depend on the *order* of the numbers p_1, p_2, \dots, p_k ; in fact, a change in the order of these numbers (i.e., a change in the columns of the probability table) is not associated with any change whatsoever in the experiment α itself. Thus, the first condition says that

E.1. The value of the function $H(p_1, p_2, \dots, p_k)$ remains invariant under any rearrangement of the numbers p_1, p_2, \dots, p_k .

The second condition is also equally natural :

E.2. The function $H(p_1, p_2, \dots, p_k)$ is continuous, i.e., it varies by a small amount for small variations in the probabilities p_1, p_2, \dots, p_k .

In fact, a small change in the probabilities must evidently correspond to a small change in the amount of uncertainty of the experiment.

The third condition we now introduce is slightly more complex. In order to have an insight into what it consists of, we first presume that the experiment α has in all three outcomes A_1, A_2, A_3 , i.e., its probability table has the form

<i>Outcomes of the experiment</i>	A_1	A_2	A_3
<i>Probabilities</i>	p_1	p_2	p_3

The measure of uncertainty $H(\alpha)$ of this experiment equals $H(p_1, p_2, p_3)$. The uncertainty arises in this case because of the fact that we do not know specifically which of the three outcomes of α will occur. We shall now clarify in two parts which of these outcomes of α actually occurs. First, we determine whether either of the first two outcomes A_1 and A_2 , or even the last outcome A_3 , has occurred; this means that our experiment α is replaced by a new experiment β with the probability table

<i>Outcomes of experiment</i>	B	A_3
<i>Probabilities</i>	$p_1 + p_2$	p_3

The measure of uncertainty of this new experiment is obviously $H(\beta) = H(p_1 + p_2, p_3)$. It is clear that the uncertainty measure of α must be greater than that of β , this is connected with the fact that knowledge of the outcome of β does not yet completely determine the outcome of α , since even after the outcome of β is revealed there may still remain some uncertainty in the outcome of α .

It is not difficult to answer the question as to exactly how much greater the uncertainty measure of α must be than that of β . Let us repeat an experiment α many times and each time reveal at the beginning whether experiment β had the outcome B or A_3 . It is then clear that in certain cases, in those in which α has the outcome A_3 , this revelation completely solves the problem of the outcome of α , too. In other cases, in those when α has outcome A_1 or A_2 , after having ascertained the outcome of β , it is appropriate to determine precisely which of these two outcomes of α occurs, which is equivalent to determining the outcome of a new experiment β' with the probability table

Outcomes of experiment	A_1	A_2
Probabilities	$\frac{p_1}{p_1 + p_2}$	$\frac{p_2}{p_1 + p_2}$

The measure of uncertainty of this experiment β' is obviously $H(\beta') = H[(p_1/(p_1 + p_2)), (p_2/(p_1 + p_2))]$. But since the probability (i.e., the average frequency) of a case in which, after the realization of β , it is further necessary to determine the outcome of β' , is equal to $p_1 + p_2$, it is natural to assume that the measure of uncertainty $H(\alpha)$ of α must exceed the measure of uncertainty $H(\beta)$ of β by the quantity $(p_1 + p_2) H(\beta')$, i.e., that the equation

$$H(p_1, p_2, p_3) = H(p_1 + p_2, p_3) + (p_1 + p_2) \times H\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right)$$

must be satisfied. The same considerations applied to an experiment α with the probability table

Outcomes of experiment	A_1	A_2	A_3	\dots	A_k
Probabilities	p_1	p_2	p_3	\dots	p_k

lead to the following third property of the function $H(p_1, p_2, \dots, p_k)$:

E.3 The function $H(p_1, p_2, \dots, p_k)$ satisfies the relation

$$H(p_1, p_2, \dots, p_k) = H(p_1 + p_2, p_3, \dots, p_k) + (p_1 + p_2) H\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right). \quad (1)$$

This relation signifies that the uncertainty $H(\beta)$ of β , with the probability table

Outcomes of experiment	B	A_3	\dots	A_k
Probabilities	$p_1 + p_2$	p_3	\dots	p_k

obtained by the identification of the first two outcomes of the experiment α , equals the uncertainty $H(\alpha)$ of α minus the measure of uncertainty of the experiment β' multiplied by $p_1 + p_2$. This seems quite natural since the experiment β' consists precisely of determining specifically which of the first two outcomes of α will occur, if one of these two outcomes is known to occur.

It can be shown that conditions *E.1* through *E.3* completely determine the form of the function $H(p_1, p_2, \dots, p_k)$: *the only function satisfying all these conditions has the form*[†]

$$H(p_1, p_2, \dots, p_k) = c(-p_1 \log p_1 - p_2 \log p_2 - \dots - p_k \log p_k). \quad (*)$$

However, the proof of this fact is not quite straightforward; it was first given by Faddeev [45]. Later, it was also shown that condition *E.2* can indeed be considerably weakened. For example, it can be replaced by the condition *E.2a*: *the function $H(p, 1-p)$ is continuous at the point $p = 0$ (i.e., $H(p, 1-p) \rightarrow H(0, 1)$ as $p \rightarrow 0$), or the condition *E.2b*: *the function $H(p, 1-p)$ does not change sign and is bounded on the interval $0 \leq p \leq 1$* ; if either of the conditions *E.2a* or *E.2b* is valid, then formula(*) also follows uniquely from conditions *E.1* and *E.3*. Some other admissible versions of weakening condition *E.2* and an extensive list of relevant references can be found, for example, in Daróczy [43]; see also Aczél, Forte and Ng [41]. However, we shall not further overstretch our treatment to the utmost generality. Following Shannon [21], we shall not only regard all three conditions *E.1*—*E.3* to be valid but we shall also supplement them with one more condition, whose validity can, in principle, be proved by using conditions *E.1*—*E.3*, but which is postulated here for the sake of considerably simplifying our reasonings*

In the sequel, an important role is played by the function $H(1/k, 1/k, \dots, 1/k)$, the measure of uncertainty of an experiment α_0 having k *equally probable* outcomes. It is obvious that, by virtue of the fact that all outcomes of α_0 are equally probable, the amount of uncertainty $H(\alpha_0)$ depends only on the number of outcomes k , i.e., it is a function of a *single argument* k : $H(1/k, 1/k, \dots, 1/k) = f(k)$. It is also clear that the amount of uncertainty of α_0 must be larger, the larger is the number k of these outcomes. Thus, we can assert that

E.4. The function $H(1/k, 1/k, \dots, 1/k) = f(k)$ increases with k .

We now show that the function $H(p_1, p_2, \dots, p_k)$ satisfying conditions *E.1*—*E.4*, necessarily has the form (*) (where c is some positive number). For this, we must slightly generalize equation (1), whose validity is guaranteed by condition *E.3*. We first show that

$$H(p_1, \dots, p_k) = H(p_1 + \dots + p_i, p_{i+1}, \dots, p_k) + (p_1 + \dots + p_i) \times H\left(\frac{p_1}{p_1 + \dots + p_i}, \frac{p_2}{p_1 + \dots + p_i}, \dots, \frac{p_i}{p_1 + \dots + p_i}\right), \quad i < k.$$

(The meaning of this equation is obviously similar to that of the original relation (1) with the only difference being that here the i outcomes A_1, A_2, \dots, A_i of experiment α are combined together as the sole outcome B of experiments β .) When $i = 2$ this equation coincides with (1) and is, consequently, valid by virtue of condition *E.3*. We now assume its validity to be proved already for some value i ; in such a case, by making use of its validity also for $i = 2$,

[†]If the coefficient c is required to be positive, then it is necessary to specify also that the function $H(p_1, p_2, \dots, p_k)$ must be non-negative (of course, it suffices to include in the basic conditions the requirement that one variable, say, $H(\frac{1}{2}, \frac{1}{2})$ be non-negative). We further note that if the basic system of logarithms is not already fixed, then the multiplier c can be discarded in formula (*) (since $c \log_a p = \log_b p$, where $b = a^{1/c}$).

This equality, fairly complex in form, expresses in most general terms the addition law of entropies enunciated in Sec. 2.2.†

The extension (2) of equation (1) will be needed by us not in its own right but in its application to the function $f(k)$. We assume that $k = lm$, where l and m are some integers, and that the $k = lm$ probabilities p_1, p_2, \dots, p_k , entering formula (2), are all equal to each other (and, consequently, equal to $1/lm$). In such a case, the left-hand side of this formula is equal to $f(lm)$. We further assume that the groups $(p_1, \dots, p_{i_1}), (p_{i_1+1}, \dots, p_{i_2}), \dots, (p_{i_s+1}, \dots, p_k)$, appearing in the equality (2), consist each of l numbers; in such a case the number of such groups is m . In addition, we have

$$p_1 + \dots + p_{i_1} = p_{i_1+1} + \dots + p_{i_2} = \dots = p_{i_s+1} + \dots + p_k = l \times \frac{1}{lm} = \frac{1}{m}.$$

Hence, the first line in the right-hand side of equality (2) reduces to $H(1/m, 1/m, \dots, 1/m) = f(m)$. Concerning the remaining members on the right-hand side of (2), the number of these members is equal to m and they are all given by

$$\begin{aligned} & (p_1 + \dots + p_{i_1}) H\left(\frac{p_1}{p_1 + \dots + p_{i_1}}, \dots, \frac{p_{i_1}}{p_1 + \dots + p_{i_1}}\right) \\ &= \frac{1}{m} H\left(\frac{1/ml}{1/m}, \dots, \frac{1/ml}{1/m}\right) = \frac{1}{m} H\left(\frac{1}{l}, \dots, \frac{1}{l}\right) = \frac{1}{m} f(l). \end{aligned}$$

Thus, in the case considered, equation (2) assumes the simple form

$$f(lm) = f(m) + m \times \frac{1}{m} f(l) = f(m) + f(l). \quad (2a)$$

From (2a) it follows in particular that

$$\begin{aligned} f(k^2) &= f(k \times k) = f(k) + f(k) = 2f(k), \\ f(k^3) &= f(k^2 \times k) = f(k^2) + f(k) = 3f(k), \\ f(k^4) &= f(k^3 \times k) = 4f(k), \end{aligned}$$

and, in general, that

$$f(k^n) = nf(k). \quad (2b)$$

We know that relation (2a) holds for the function $f(k) = c \log k$. It is also routine to show that $c \log k$ is the *only* function that satisfies relation (2a) and condition E.4. In fact, suppose that k and l are two arbitrary positive integers. Choose some other large integer N and find a number n such that

$$l^n \leq k^N < l^{n+1}.$$

By E.4,

$$f(l^n) \leq f(k^N) < f(l^{n+1}),$$

†It is trivial to be convinced in that, if $i_1 = l, i_2 = 2l, i_3 = 3l, \dots, k = (s+1)l$ and the variables $p_1, p_2, \dots, p_{i_1}; p_{i_1+1}, p_{i_1+2}, \dots, p_{i_2}; \dots$ are the probabilities of outcomes $A_1B_1, A_1B_2, \dots, A_1B_{i_1}; A_2B_1, A_2B_2, \dots, A_2B_{i_2}; \dots$ of a compound experiment $\alpha\beta$ (the sums $p_1 + p_2 + \dots + p_{i_1}, p_{i_1+1} + p_{i_1+2} + \dots + p_{i_2}, \dots$ are equal in such a case to the probabilities of outcomes A_1, A_2, \dots of an experiment α), then the equation (2) turns into the addition law of entropies.

or, by virtue of (2b),

$$nf(l) \leq Nf(k) < (n+1)f(l).$$

This implies that

$$\frac{n}{N} \leq \frac{f(k)}{f(l)} \leq \frac{n+1}{N}.$$

We now note that from $l^n \leq k^N < l^{n+1}$ it follows that

$$n \log l \leq N \log k < (n+1) \log l,$$

or

$$\frac{n}{N} \leq \frac{\log k}{\log l} < \frac{n+1}{N}.$$

Thus, the ratios $f(k)/f(l)$ and $\log k/\log l$ lie within one and the same narrow bounds and, consequently, must be close to each other :

$$\left| \frac{f(k)}{f(l)} - \frac{\log k}{\log l} \right| < \frac{1}{N}.$$

But since the last inequality holds for *every* value N , it follows that

$$\frac{f(k)}{f(l)} = \frac{\log k}{\log l},$$

or

$$\frac{f(k)}{\log k} = \frac{f(l)}{\log l}.$$

This relation holds for *each* of the two numbers k and l ; consequently,

$$\frac{f(k)}{\log k} = \frac{f(l)}{\log l} = c,$$

where c does not depend on k and l , and hence,

$$f(k) = c \log k.$$

But since $f(k)$ is an increasing function, therefore $c > 0$.

We now suppose that p_1, p_2, \dots, p_k are arbitrary fractions

$$p_1 = \frac{q_1}{p}, \quad p_2 = \frac{q_2}{p}, \quad \dots, \quad p_k = \frac{q_k}{p}$$

(q_1, \dots, q_k, p being integers and p being the common denominator of all these fractions), such that all of them are less than unity and $p_1 + p_2 + \dots + p_k = 1$. According to formula

(2) (p. 97), we have

$$\begin{aligned}
 f(p) &= H\left(\underbrace{\frac{1}{p}, \frac{1}{p}, \dots, \frac{1}{p}}_{p \text{ times}}\right) \\
 &= H\left(\underbrace{\frac{1}{p}, \frac{1}{p}, \dots, \frac{1}{p}}_{q_1 \text{ times}}, \underbrace{\frac{1}{p}, \frac{1}{p}, \dots, \frac{1}{p}}_{q_2 \text{ times}}, \dots, \underbrace{\frac{1}{p}, \frac{1}{p}, \dots, \frac{1}{p}}_{q_k \text{ times}}\right) \\
 &= H\left(\frac{q_1}{p}, \frac{q_2}{p}, \dots, \frac{q_k}{p}\right) + \frac{q_1}{p} H\left(\underbrace{\frac{1}{q_1}, \frac{1}{q_1}, \dots, \frac{1}{q_1}}_{q_1 \text{ times}}\right) \\
 &\quad + \frac{q_2}{p} H\left(\underbrace{\frac{1}{q_2}, \frac{1}{q_2}, \dots, \frac{1}{q_2}}_{q_2 \text{ times}}\right) + \dots + \frac{q_k}{p} H\left(\underbrace{\frac{1}{q_k}, \frac{1}{q_k}, \dots, \frac{1}{q_k}}_{q_k \text{ times}}\right) \\
 &= H(p_1, p_2, \dots, p_k) + p_1 f(q_1) + p_2 f(q_2) + \dots + p_k f(q_k).
 \end{aligned}$$

This implies that

$$\begin{aligned}
 H(p_1, p_2, \dots, p_k) &= f(p) - p_1 f(q_1) - p_2 f(q_2) - \dots - p_k f(q_k) \\
 &= (p_1 + p_2 + \dots + p_k) f(p) - p_1 f(q_1) - p_2 f(q_2) - \dots - p_k f(q_k) \\
 &= p_1(f(p) - f(q_1)) + p_2(f(p) - f(q_2)) + \dots + p_k(f(p) - f(q_k)).
 \end{aligned}$$

But since

$$f(p) - f(q_1) = c \log p - c \log q_1 = -c \log \frac{q_1}{p} = -c \log p_1,$$

$$f(p) - f(q_2) = -c \log p_2, \dots, f(p) - f(q_k) = -c \log p_k,$$

we finally obtain

$$H(p_1, p_2, \dots, p_k) = c(-p_1 \log p_1 - p_2 \log p_2 - \dots - p_k \log p_k).$$

The last equality has so far been proved only for *rational* values p_1, p_2, \dots, p_k . But by the continuity of the function $H(p_1, p_2, \dots, p_k)$ it follows that it is true for *every* p_1, p_2, \dots, p_k . This completes our proof.

3

The Solution of Certain Logical Problems by Calculating Information

3.1. Simple examples

In order to illustrate the practical versatility of the concepts and propositions of Chapter 2, we analyze here some amusing problems of the sort collected by Kordemskii [59]. In Sections 1 and 2 we shall formulate some specific examples of such problems and here we shall frequently use heuristic arguments based on the intuitive notion of information. A deeper and more rigorous discussion of the reasonings in these sections will be postponed to the concluding Section 3 of this chapter.

We start with the well-known logical problem concerning a ‘town of liars and a town of non-liars,’ which is quite popular among mathematics enthusiasts in high schools.

Problem 23. *Suppose we know that the inhabitants of a certain town A always tell the truth, while those of a neighbouring town B always lie. An observer O knows that he is in one of these two towns but does not know specifically which one. By interrogating a person he encounters O must determine the town he is in, or the town in which his collocutor resides (residents of A can visit B and vice versa), or both facts together. What is then the least number of questions O must ask (the collocutor is to answer only ‘yes’ or ‘no’ to all questions asked by O)?*

Suppose that O must determine the town he is in. Here the experiment β , whose result is of interest to him, can have two outcomes (this experiment consists of finding out in which of the *two* towns, A or B , the observer O is). If we assume that O has no information beforehand as to which of the two towns he is in, then these two outcomes should be considered as equally possible and consequently, the entropy $H(\beta)$ of β (i.e., the ‘total’ amount of information contained in the outcome of this experiment) equals one bit. Furthermore, the experiment α , in which O puts one question to the collocutor, can also have two outcomes (the latter may answer ‘yes’ or ‘no’); hence the entropy $H(\alpha)$ of this experiment (i.e., the ‘total’ amount of information contained in the answer to the question asked) is at most equal to one bit. The question that now arises is whether

experiment α can be so set up that the information $I(\alpha, \beta)$ contained in α about experiment β equals the entropy $H(\beta) = 1$ of β , i.e., that *the outcome of α completely determines the outcome of β* . Let us now recall that the sole relationship between the information $I(\alpha, \beta)$ and entropy $H(\alpha)$ consists of the facts that

$$I(\alpha, \beta) \leq H(\alpha) \quad (\text{since } I(\alpha, \beta) = H(\alpha) - H_{\beta}(\alpha)).$$

Since $H(\alpha)$ can equal 1, we can expect in general that, subject to a successful choice of experiment (i.e., the question) α , the equality

$$I(\alpha, \beta) = H(\beta)$$

may hold. For this, the only requirements are that the question pertaining to experiment α be such that an affirmative or negative answer to it is equally probable† (it is only in this case that the equality $H(\alpha) = 1 = H(\beta)$ holds), and that the outcome of experiment β determines that of α (it is only subject to this condition that the equality $I(\alpha, \beta) = H(\alpha)$, or $H_{\beta}(\alpha) = 0$, holds, indicating that the question pertaining to experiment α is ‘directed straight’ to ascertaining the outcome of β and an answer to this question contains no ‘extraneous’ information). All these restrictions are satisfied by the question ‘*Do you live in this town?*’, which completely solves the problem. (A positive answer to this question can be given only in town A and a negative answer only in town B .)

It is quite obvious that O can ascertain the town in which his collocutor resides by asking a single question: for this it suffices to put any question, whose answer is known to O beforehand (say, ‘*Am I in a town?*’, or ‘*Does 2×2 make four?*’).

If, however, O has to know both the town he is in and the town in which his collocutor resides, then he is called upon to determine the outcome of the joint experiment $\beta_1\beta_2$, where β_1 consists of determining where O is and β_2 the place of residence of his collocutor. The entropy $H(\beta_1\beta_2)$ of this experiment is greater than the entropy $H(\beta_1)$ of β_1 : $H(\beta_1\beta_2) = H(\beta_1) + H_{\beta_1}(\beta_2)$ (see Sec. 2 of Chapter 2). In other words, in this case the information required is greater than 1 bit (recall that $H(\beta_1) = 1$). Since the entropy $H(\alpha)$ of an experiment α (which consists of asking a question) with two outcomes cannot exceed 1, a single experiment α does not provide an opportunity to obtain information equal to $H(\beta_1\beta_2)$, i.e., does not enable us to determine completely the outcome of $\beta_1\beta_2$ (except for the completely uninteresting case in which the conditional entropy $H_{\beta_1}(\beta_2)$ is 0, i.e., in which the outcome of β_1 determines the outcome of β_2 ; such is the situation when the residents of A cannot enter B and conversely).

†Subject to the condition that O be in either A or B and that his collocutor does reside in either A or B are equally probable.

Thus, an estimate of the amount of information yields us a rigorous proof of the fact that a single question (no matter how it is put !) does not enable us to determine directly both the town in which O is and the town in which his collocutor resides. If, however, O puts two questions (i.e., carries out a joint experiment $\alpha_1\alpha_2$, having four possible outcomes), then he can indeed ascertain the outcome of experiment $\beta_1\beta_2$ (the outcome of β_1 can be determined with the aid of the question pertaining to experiment α_1 , and that of β_2 by the question pertaining to experiment α_2).

In the next problem, some of the hypotheses of Problem 23 bear a more complex character.

Problem 24. *Suppose that there are three towns A , B and C . The inhabitants of A always tell the truth, those of B only tell lies and those of C alternately tell the truth and lie. An observer O desires to find out the town in which he is and the town in which a person he encounters resides. How many questions need he put to his collocutor if all the questions are to have only 'yes' or 'no' answer?*

Here we must determine which of the *nine* possible outcomes of experiment β is realized (O may be in any one of the three towns A , B and C and, independent of this, his collocutor may reside in any one of the same three towns). If we assume that O has no prior information about experiment β , then all these nine outcomes can be considered to be equally probable and the entropy $H(\beta)$ of β (and, consequently, also the amount of information obtained by ascertaining the outcome of β) equals $\log 9$. Suppose that the joint experiment $A_k = \alpha_1\alpha_2 \dots \alpha_k$ consists of having O ask k questions. Since he may receive an affirmative or a negative answer to each question, the entropy of each experiment $\alpha_1, \alpha_2, \dots, \alpha_k$ does not exceed one bit. On the other hand,

$$H(\alpha_1\alpha_2) = H(\alpha_1) + H_{\alpha_1}(\alpha_2) \leq H(\alpha_1) + H(\alpha_2)$$

(because $H_{\alpha_1}(\alpha_2) \leq H(\alpha_2)$) and similarly,

$$H(A_k) = H(\alpha_1\alpha_2 \dots \alpha_k) \leq H(\alpha_1) + H(\alpha_2) + \dots + H(\alpha_k) \leq k$$

(a rigorous proof of this inequality is easy to obtain by mathematical induction). This can be verbalized differently as follows: If the answer to each question yields us information not exceeding one bit, then by asking k questions we can obtain information not greater than k bits. Hence if $k = 3$, then the information given to us is less than $\log 9$ (it can at most equal $3 = \log 8 < \log 9$) and, thus, three questions will not ensure that we can always determine both the place in which O is and the place in which his collocutor resides. However, four adroitly put questions may possibly do the trick (because it can then be asserted that $H(A_4) \leq 4 = \log 16$). Indeed, it is easy to see that the following four questions do assure the revelation of all that is of interest to O :

- (i) Do I happen to be in one of the towns A and B ?
- (ii) Do I happen to be in town C ?
- (iii) Do you reside in town C ?
- (iv) Do I happen to be in town A ?

In fact 'yes' or 'no' answers to *both* the questions (i) and (ii) immediately indicate that the collocutor of O resides in C . Suppose, for instance, that the answers to both of these questions are in the affirmative (the case in which both answers are negative is analyzed similarly). In this case, a negative (obviously incorrect) answer to question (iii) implies that the answer to question (ii) is correct and eliminates the necessity of further asking question (iv); a positive (correct) answer to question (iii) means that the answer to question (i) is trustworthy and in order to find out the town in which O is, it is necessary to put question (iv) (the answer to which is known to be incorrect). An affirmative answer to (i) and a negative answer to (ii), as well as the converse position, indicate that the collocutor of O resides in A or B . In such a case, a negative (i.e., correct) answer to question (iii) shows that the respondent resides in A and question (iv) is needed only if the answer to (ii) is negative; a positive (incorrect) answer to question (iii) means that the collocutor of O resides in B and question (iv) is found necessary only if the answer to (ii) is positive.

The following is one more example of a similar sort of problem (see Problem 283 in [59]).

Problem 25. *How many questions are necessary to determine a positive integer thought of by a collocutor assuming that the conceived integer does not exceed 10 (or 100, or 1000 or an arbitrary positive integer n) and only 'yes' or 'no' can be given as answers to all questions ?*

Suppose we know that the thought of number does not exceed 10. In such a case, an experiment β , consisting of the determination of this number, can have 10 different outcomes. Until the first question is put and answered, we can consider all these outcomes to be equally probable so that the entropy $H(\beta)$ of β (i.e., the requisite information) equals $\log 10 \approx 3.32$ bits. We consider a joint experiment $A_k = \alpha_1 \alpha_2 \dots \alpha_k$ in which k questions are asked. The entropy of α_1 , where α_1 consists of asking a single question, does not exceed 1 bit since α_1 can have only two outcomes (positive and negative answers to the question); hence the entropy of A_k does not exceed k bits (see p. 103). On the other hand, the information $I(A_k, \beta)$ concerning experiment β that is contained in the joint experiment A_k cannot exceed the total information contained in the outcome of A_k , i.e., the entropy $H(A_k)$. In order that the outcome of A_k completely determine the outcome of β , it is necessary that the equality $I(A_k, \beta) = H(\beta)$ hold. Hence, we conclude in this case that

$$\log 10 = H(\beta) = I(A_k, \beta) \leq H(A_k) \leq k,$$

i.e.,

$$k \geq \log 10 \approx 3.32,$$

and since k is an integer,

$$k \geq 4.$$

Let us now show that by asking just four questions the outcome of β can indeed be completely determined, i.e., we can find the number x that was thought of. It is easy to visualize the procedure to follow for this purpose. In the first place, it is natural to put the first question in such a way that the information contained in the answer to it, that is, the entropy $H(\alpha_1)$, is the maximum possible. In other words, the information actually equals one bit. For this, it is necessary that both outcomes of our experiment α_1 be equally probable. The further requirement is that the information $I(\alpha_1, \beta)$ about β , contained in α_1 , be equal to but not less than the entropy $H(\beta)$ of β . This demands that the answer to the first question contain no 'extraneous' information, i.e., that the conditional entropy $H_{\beta}(\alpha_1)$ be zero (in other words, the outcome of α_1 is fully determined by the outcome of β). These considerations clearly prescribe how the first question ought to be put. We partition the set of all possible values of x (i.e., the set of positive integers from 1 to 10) into two *numerically equal* parts (since the two outcomes of α_1 must be equally probable) and then ask to which of these two parts x belongs. Thus, we may ask, say, if x is greater than 5. In this case, obviously,

$$I(\alpha_1, \beta) = H(\beta) - H_{\alpha_1}(\beta) = 1,$$

i.e.,

$$H_{\alpha_1}(\beta) = p(A_1) H_{A_1}(\beta) + p(A_2) H_{A_2}(\beta) = H(\beta) - 1$$

(A_1 and A_2 are the two outcomes of α_1 ; $p(A_1) = p(A_2) = \frac{1}{2}$); in addition,

$$H_{A_1}(\beta) = H_{A_2}(\beta) = H(\beta) - 1,$$

so that for *every* outcome of α_1 , the entropy of the experiment β we are interested in, decreases by 1 bit. Furthermore, in exactly the same manner, we divide the new set of permissible values of x into *two equal* (or, at least, *approximately equal*) parts, and determine to which of them x belongs (if x is greater than 5, then ask whether this number is larger than 7; if, however, x does not exceed 5, then question whether x is larger than 3), and so on. Each time, by partitioning the set of admissible values of x into two parts, *as numerically equal as possible*,

we can certainly determine x by asking only four questions.†

In exactly the same way, we can show that the least number k of questions enabling us to determine an unknown number x , which may have 100 or 1000 values, is given by the inequality $k \geq \log 100 \approx 6.64$ and, correspondingly, by $k \geq \log 1000 \approx 9.97$. Since in all such cases k is an integer, this implies that

$$k \geq 7, \text{ or (correspondingly) } k \geq 10.$$

In general, the least number k of questions enabling one to find an unknown number x having one of n admissible values is given by the inequalities

$$k - 1 < \log n \leq k \quad (\text{or } 2^{k-1} < n \leq 2^k). \quad (1)$$

We note also that

$$k \geq \log n,$$

in all the cases; moreover, $k = \log n$ if and only if the number n is an integral power of 2 and, consequently, $\log n$ is an integer. However, when n is very large, the difference between the numbers k and $\log n$ is found to be quite small in comparison to these numbers themselves (because, for large n , the quantity $\log n$ is also large and the difference $k - \log n$ does not always exceed unity). Thus, we can assume that for large n , the ratio of $\log n$ (the entropy of β under consideration) to the information (1 bit) about β contained in experiment α (which consists of finding the answer to a single question), quite precisely indicates the number k of experiments that are involved to determine the outcome of β .

At first sight, Problem 25 appears to be as artificial as its two predecessors; we shall see later, however, that it has serious engineering applications.†† A more detailed discussion of the solution of this problem (including also a more general formulation of its conditions) is deferred to Sec. 3 of this chapter.

The next problem is quite similar to Problem 25.

Problem 26. *A person thinks of two (distinct) numbers not exceeding 100. How many questions are necessary to find these numbers if each question can have only 'yes' or 'no' answers?*

†Obviously, after we find that the number x has one of m values, where m is odd (say, $m = 5$), we cannot secure strictly equally probable outcomes of the succeeding experiment α_{i+1} because m possible values of x are impossible to split into numerically equal parts. Hence the entropy $H(\alpha_{i+1})$ of experiment α_{i+1} will be less than 1. This implies that our questioning will not be most profitable from the viewpoint of information obtained, i.e., that by using the same number of questions, we can find an unknown number even when the set of its possible values is a larger number (thus, using four questions we can find an unknown number which is not just one of 10 but even one of $2^4 = 16$ possible values).

††It should nevertheless be indicated that in spite of the recreative formation of Problems 23–24, a sufficiently serious meaning lies concealed in them (see pp. 121–123).

In this case, experiment β whose outcome must be determined, can have $\binom{100}{2} = 4950$ different outcomes. If, as usual, we consider all these outcomes to be equally probable, then the entropy $H(\beta)$ of β (i.e., the information that we obtain after having determined the outcome of β) equals $\log 4950$. But, since the information that can be provided by an answer to a single question does not exceed one bit (because experiment α , which consists of asking a single question can have but two outcomes 'yes' or 'no'), the least number of questions we must ask to be able to always determine the outcome of β can never be less than $\log 4950 \approx 12.27$ (cf. the solution of Problem 25). Thus, if we ask less than thirteen questions, it is certainly possible that we shall not succeed in determining both of the unknown numbers.

It is also easy to see that thirteen adroitly put questions always enable us to find the two numbers. In order to achieve this, it is necessary that the information $I(\alpha, \beta)$ obtained concerning the outcome of experiment β contained in the outcome of experiment α in which a single question is asked (more precisely, in which each of the questions is asked), be as close to one bit as possible. Hence, it is plain that questions are necessarily so put that the answers 'yes' and 'no' are equiprobable or nearly equiprobable. And for this purpose, it suffices that to begin with, we partition the 4950 outcomes of β into two numerically equal parts (such that each part contains 2475 outcomes) and determine to which of these parts the real outcome of β belongs (i.e., we should ask in the first place whether or not the two unknown numbers belong to the group containing the first 2475 pairs of numbers). Next, in exactly the same way, it is necessary to divide into two numerically equal parts (as far as possible) that group of outcomes to which the outcome of our interest belongs, and then determine to which of these two smaller parts the two unknown numbers belong, and so on. It is clear that here we invariably determine the pair of unknown numbers with the aid of not more than thirteen questions.

We further remark that the distinction between Problems 26 and 25 can be considered to be purely verbal. It is clear that in solving Problem 25 a role is played only by the *total number* n of those numbers, one of which is the number thought of. In addition, obviously, we can always consider these n numbers to be indexes of arbitrarily chosen objects, say, n indexes of n cars, or n pairs of numbers, or n given arbitrary groups of numbers, and so on—this has no influence on the solution of the problem. However, if we consider that n in Problem 25 equals 4950 and that the considered 4950 numbers index that set of all possible pairs of numbers, each of which does not exceed 100, then we arrive at Problem 26.

In exactly the same way, we can show that the minimum number of questions we need ask to determine the m conceived numbers, not exceeding n , equals the least integer k such that $k \geq \log \binom{n}{m}$. If, however, it is known, say, that either one number not exceeding n is thought of, or no number is thought of, then in

order to find out whether a number has been thought of and if so, precisely what number, it is necessary to put questions not less than $\log(n+1)$ and not more than $\log(n+1)+1$. In fact, in this case, the number of possible outcomes of the corresponding experiment β is $n+1$ (the unity in this sum corresponds to the case where no number is thought of). Finally, if we assume that *not more* than m numbers are thought of, where $m \leq n/2$, each of which does not exceed n , then the number of questions necessary to determine how many and exactly what numbers were thought of lies between

$$\log \left[\binom{n}{m} + \binom{n}{m-1} + \dots + \binom{n}{1} + 1 \right],$$

and

$$\log \left[\binom{n}{m} + \binom{n}{m-1} + \dots + \binom{n}{1} + 1 \right] + 1.$$

In fact, the experiment β considered here can have $\binom{n}{m} + \binom{n}{m-1} + \dots + \binom{n}{1} + 1$ different outcomes (because what was thought of may turn out to be a group in the $\binom{n}{m}$ groups of m numbers, or one of the $\binom{n}{m-1}$ groups of $m-1$ numbers, \dots , or one of $\binom{n}{1} = n$ individual numbers, or even no number at all). Renumbering these $N = \binom{n}{m} + \binom{n}{m-1} + \dots + \binom{n}{1} + 1$ outcomes of experiment β as the numbers from 1 to N , we arrive at Problem 25 (where the number n has been replaced by N). Later, we shall make further use of this remark.

3.2. The counterfeit coin problem

The starting point of this section is the following problem, closely allied to Problem 25.

Problem 27. *There are 25 coins of the same denomination. Of these, 24 are of identical weight and the one counterfeit coin is slightly lighter than the others. The question is how many weighings on a beam balance are necessary to find the counterfeit coin ? (See Problem 277, 1 and 2 in [59].)*

The experiment β whose result must be determined has in this case 25 possible outcomes (any of the 25 coins may turn out to be counterfeit). It is natural to suppose all these outcomes to be equally probable so that $H(\beta) = \log 25$. In other words, the determination of a counterfeit coin is related in the given case to obtaining the information measured by the number $\log 25$. The experiment α_1 , consisting of one (arbitrary) weighing, can have three outcomes (the left or the right beam may be lighter or both may be equal); hence, $H(\alpha_1) \leq \log 3$ and

the information $I(\alpha_1, \beta)$ obtained from such an experiment does not exceed $\log 3$. We now consider the joint experiment $A_k = \alpha_1 \alpha_2 \dots \alpha_k$ consisting of k consecutive weighings; it gives information not exceeding $k \log 3$ (see p. 103). If experiment A_k enables us to determine completely the outcome of the experiment β , then we must have

$$H(A_k) \geq I(A_k, \beta) \geq H(\beta), \text{ or } k \log 3 \geq \log 25.$$

Hence, we infer that $3^k \geq 25$, i.e.,

$$k \geq \log_3 25 = \frac{\log 25}{\log 3},$$

and since k is an integer, we must have

$$k \geq 3.$$

It is easy to show that with the aid of three weighings the counterfeit coin can be found. If we want to gain the maximum possible information from experiment α_1 , it is necessary that the outcomes of this experiment be (as far as possible) equally probable. Suppose that m coins are placed on each beam (clearly it makes no sense to put different numbers of coins on two beams: in such a case the outcome of the corresponding experiment is known beforehand, and the information obtained is 0); the number of coins not placed on the balance is equal to $25 - 2m$. Since the probability that the counterfeit coin will turn up in a given group of n coins is $n/25$ (because all outcomes of experiment β are considered equally probable!), the three outcomes of experiment α_1 have the probabilities $m/25$, $m/25$ and $(25 - 2m)/25$. These probabilities are closest to one another when $m = 8$ and $25 - 2m = 9$. If 8 coins are placed on each beam, the first weighing (experiment α_1) allows us to select a second group of 9 coins (if the beams are equal) or 8 coins (if one of the beams is lighter), one of which is counterfeit. In both cases, in order to obtain the maximum information from the second weighing (experiment α_2) it is necessary to place three coins from this group on each of the two beams of the balance; in such case the joint experiment $\alpha_1 \alpha_2$ permits us to select a group of 3 (or of 2) coins, one of which is counterfeit. In the third weighing (experiment α_3), we place one of the remaining suspect coins on each of the two beams of the balance and easily find the counterfeit coin.

In exactly the same way, we can show that *the least number k of weighings, that enable us to determine a single counterfeit (lighter!) coin contained in a group of n coins, is given by the inequalities:*

$$3^{k-1} < n \leq 3^k, \text{ or } k-1 < \frac{\log n}{\log 3} \leq k. \quad (2)$$

If n is large, then this number k is given, with a sufficient accuracy, by the ratio $\log n / \log 3$, i.e., by the ratio of the entropy of experiment β , consisting of the determination of the counterfeit coin, to the maximum information which can be obtained in a single weighing (see p. 106).

In the following, we shall use a similar conclusion related to an even more general setting of the problem. In the first place, it is clear that if we have n coins with one counterfeit among them and we know that the counterfeit coin is slightly heavier than the others, then the least number of weighings k on a beam balance that enables us to detect this counterfeit coin, is given by the same inequalities (2): in practice, the substitution of a heavier coin for the lighter one does not alter our arguments. We now consider a more general case in which our n coins are divided into two groups; group A , containing a coins and group B , containing $b = n - a$ coins, it being known that one of these n coins is counterfeit and that if this coin belongs to group A (resp. B), then it is lighter (resp. heavier) than the rest, and show that here also the least number of weighings k that enables us to find the counterfeit coin is given by inequalities (2).† For $b = 0$, this statement reverts to the previous case.

In fact, it is clear that the experiment β in which we are interested can obviously have n different outcomes. Hence $3^k \geq n$; otherwise, the experiment $A_k = \alpha_1 \alpha_2 \dots \alpha_k$, consisting of k subsequent weighings, can in no way uniquely determine the outcome of β (because in this case $I(A_k, \beta) \leq H(A_k) \leq k \log 3 = \log 3^k < \log n = H(\beta)$; the outcomes of β are considered, as usual, to be equally probable). On the other hand, when $n \leq 3^k$ the counterfeit coin can always be separated out by k weighings; this is easy to show by using, say, mathematical induction. In fact, if $k = 1$, i.e., $n = 1, 2$ or 3 , then our assertion is almost obvious (with the one exception indicated in the preceding footnote): for $n = 1$ the counterfeit coin is known always, but for $n = 2$ (and $a = 2$ or $b = 2$) and for $n = 3$, it suffices to compare the weights of two coins from one group in order to determine the counterfeit one. We now suppose it to be already known that for $n \leq 3^k$ the counterfeit coin can always be separated out with the aid of not more than k weighings. Let $3^k < n \leq 3^{k+1}$. It is easy to see that in this case we can always select an even number $2x$ of coins from group A and an even number $2y$ of coins from group B such that the numbers x and y satisfy the conditions

$$2x + 2y \leq 2 \times 3^k, \quad n - (2x + 2y) \leq 3^k,$$

i.e.,

$$3^k \geq x + y \geq \frac{n - 3^k}{2}.$$

†This statement has one obvious exception: if $n = 2$, $a = b = 1$, then it is obviously quite impossible to separate out the counterfeit coin.

We now place x coins from group A and y coins from group B on each beam. Then, the number of coins not placed on the balance is $n_1 = n - 2x - 2y \leq 3^k$. If the beams balance for this weighing (experiment α_1), then we infer that the counterfeit coin is among the n_1 coins not on the balance, i.e., among the $a_1 = a - 2x$ (resp. $b_1 = b - 2y$) coins from group A (resp. B) not involved in the first weighing. If one of the beams is lighter, then one of the x coins from group A lying on the lighter pan or of the y coins from group B lying on the heavier pan is counterfeit. However, since $n_1 \leq 3^k$ and $x + y \leq 3^k$, by the assumption made we are able to separate out the counterfeit coin in both the cases by not more than k weighings.[†] Consequently, from our $n \leq 3^{k+1}$ coins, we can certainly find the counterfeit coin by making not more than $k + 1$ weighings. This conclusion also completes the proof of the statement made above.

We now consider the following problem, which is slightly more complicated.

Problem 28. *There are 12 coins of the same denomination, of which 11 have identical weight and the remaining one is counterfeit, having a weight different from all the rest (it being unknown whether it is lighter or heavier than the genuine ones). What is the least number of weighings on a beam balance that will enable us to find the counterfeit coin and determine whether it is lighter or heavier than the rest of the coins? Solve the same problem also for the case of 13 coins (see Problem 277 (3) in [59] or Problem 6(a) in [62]).*

We consider here an experiment β having 24 or 26 possible outcomes (any one among the existing 12 or 13 coins may be counterfeit, and this coin may be either lighter or heavier than the genuine coins). If all these outcomes are considered to be equally probable, the entropy $H(\beta)$ of β equals $\log 24$ or $\log 26$. Thus we are required to obtain $\log 24$, or correspondingly $\log 26$, units of information. Since from the joint experiment $A_k = \alpha_1 \alpha_2 \dots \alpha_k$, consisting of k weighings, we can obtain information not greater than $k \log 3 = \log 3^k$, and $\log 3^3 = 27$, at the first sight it seems plausible that in the case of 12 or 13 coins, three weighings will enable us to find the counterfeit coin and also to decide whether it is lighter or heavier than others. In reality, however, in the case of 13 coins three weighings may be found to be insufficient; this fact is quite simple to show with the aid of a somewhat more careful evaluation of the information obtained from the first weighing.

In fact, the first weighing may consist of placing 1, 2, 3, 4, 5 or 6 coins on each beam. We denote the corresponding experiments by $\alpha_1^{(i)}$, where i can equal

If $n > 2$, then the case in which $x = y = 1$, or $a_1 = b_1 = 1$ no longer constitutes an exception. In fact, apart from one suspect coin from group A and one from group B , we now have a certain number of coins that are known to be genuine; by comparing the weight of one of them with that of one from the suspect coins we shall be able to find the counterfeit coin from one weighing.

1, 2, 3, 4, 5 or 6. If i equals 1, 2, 3, or 4 and, as a consequence of the first weighing, the beams remain balanced, then experiment $\alpha_1^{(i)}$ indicates that one of the $13 - 2i$ coins not on the balance is counterfeit. Since this number is not less than 5, 10 (or still more) different outcomes remain possible after the first weighing. Therefore, the two succeeding weighings may not guarantee the detection of the counterfeit coin and the clarification of whether it is lighter or heavier than the rest (because $2 \log 3 = \log 9 < 10$). If i equals 5 or 6 and, in experiment $\alpha_1^{(i)}$, one beam (say, the right one) is heavier, then $\alpha_1^{(i)}$ indicates that either one of the i coins on the 'right' beam is counterfeit and heavier, or one of the i coins on the 'left' beam is counterfeit and lighter. Thus, here also, we are still left with $i + i = 2i \geq 10$ possible outcomes of experiment β , and again two weighings are insufficient to ascertain which outcome actually holds.

We now pass on to the case of 12 coins. Suppose that, in the first weighing, we place i coins on each beam (experiment $\alpha_1^{(i)}$). If, in this, the beams remain balanced (outcome E of experiment $\alpha_1^{(i)}$; we shall use similar notation in what follows), then one of the $12 - 2i$ coins not on the balance is counterfeit, which corresponds to $2(12 - 2i)$ equally probable outcomes of the experiment β under consideration (from the total number of 24 outcomes). If the right beam is heavier (outcome R), then either one of the i coins on the right beam is counterfeit and heavier, or one of the i coins on the left beam is counterfeit and lighter—these cases correspond to the $2i$ outcomes of β —in exactly the same manner, the case in which the left beam is heavier (outcome L) also corresponds to the $2i$ outcomes of β . Thus, the three outcomes of the experiment $\alpha_1^{(i)}$ have the probabilities

$$\frac{2(12 - 2i)}{24} = \frac{6 - i}{6}, \quad \frac{2i}{24} = \frac{i}{12} \quad \text{and} \quad \frac{i}{12}.$$

Hence, it immediately follows that of the six experiments $\alpha_1^{(1)}, \alpha_1^{(2)}, \dots, \alpha_1^{(6)}$, experiment $\alpha_1^{(4)}$ whose three outcomes are equally probable, has the largest entropy. Thus, the experiment $\alpha_1^{(4)}$ gives us the maximum information and it is most expedient to start with it. We now consider the two cases separately.

Case I. *The beams are balanced for the first weighing.* In this case, one of the four coins not on the balance is counterfeit. By means of two weighings we must find out which of these coins is counterfeit and also ascertain whether it is lighter or heavier than the others. Since we are left with $2 \times 4 = 8$ possible outcomes of experiment β and $2 \log 3 = \log 9 > \log 8$, we may expect that this is possible. However, if just one of our four suspect coins is placed on each beam so that two coins are not on the balance (experiment $\alpha_2^{(1)}$) and the beams remain equal, then from the next weighing we must determine specifically which of the *four* outcomes that still remain possible occurs. This is clearly impossible

to do (since $4 > 3$). If, however, we place on each beam a pair of our four suspect coins (experiment $\alpha_2^{(2)}$) and one of the two beams is heavier, then we are again left with four still possible outcomes of experiment β and have to resort again to at least two further weighings in order to determine completely which of them occurs. This gives an impression that, in the case of 12 coins also, three weighings are insufficient to solve the problem.

This inference is however premature. In fact, we have not taken into account that, after the first weighing, we have at our disposal $4 + 4 = 8$ a fortiori *genuine* coins that can participate in the second weighing. Hence, we have considerably more than two possible variants of experiment α_2 . Let us denote by $\sigma_2^{(i,j)}$ an experiment in which we place on the right beam i of our four suspect coins and on the left beam $j \leq i$ of them as well as $i - j$ definitely genuine coins (obviously, it makes no sense to place the genuine coins on both the beams). In such a case $\alpha_2^{(1,1)}$ and $\alpha_2^{(2,2)}$ are those experiments $\alpha_2^{(1)}$ and $\alpha_2^{(2)}$ that we considered above. We denote by $p(R)$, $p(L)$ and $p(E)$, respectively, the probabilities that in experiment $\sigma_2^{(i,j)}$ the right beam is heavier, the left beam is heavier, or both are equal. These probabilities are easy to calculate; they equal the ratio of the number of those outcomes of β for which $\alpha_2^{(i,j)}$ has outcome R , correspondingly, L or E , to the total number of remaining possible outcomes of β (this number is 8). Since $i + j \leq 4$, obviously all experiments $\sigma_2^{(i,j)}$ are easy to enumerate; the values of the probabilities $p(R)$, $p(L)$ and $p(E)$ corresponding to them are listed in the table on p. 114. In the last column of the table, we have given also the entropy (in bits) $H(\alpha_2^{(i,j)})$ of experiments $\sigma_2^{(i,j)}$, which equals $-p(E) \log p(E) - p(R) \log p(R) - p(L) \log p(L)$.

From this table, we see that experiments $\alpha_2^{(2,1)}$ and $\alpha_2^{(3,0)}$ have the largest entropy. Hence, in order to gain the maximum information, it is necessary in the process of the second weighing either to put two of the four suspect coins on one beam and one of suspect coins and one definitely genuine coin on the other beam, or to put three suspect coins on one beam and three definitely genuine ones on the second beam. It is easy to see that, in both cases, we can then completely determine, by the third weighing, the outcome of β . Indeed, if experiment $\alpha_2^{(2,1)}$ or $\alpha_2^{(3,0)}$ has outcome E , then the only suspect coin not on the balance in the second weighing is counterfeit; in order to find out also whether it is lighter or heavier than the others, it is necessary to compare its weight with that of one of 11 definitely genuine coins (this is the third weighing). If experiment $\alpha_2^{(2,1)}$ has outcome R , then either one of the two coins on the right beam is counterfeit and is heavier than the others, or the lone suspect coin on the left beam is counterfeit and it is lighter than the genuine ones. Comparing the weights of the two coins on the right beam (by a third weighing) we are able to know the outcome of β (if these coins have the same weight, then the third of

i	j	$p(E)$	$p(R)$	$p(L)$	$H(\alpha_2^{(i,j)})$
1	1	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	1.50
1	0	$\frac{3}{4}$	$\frac{1}{8}$	$\frac{1}{8}$	1.06
2	2	0	$\frac{1}{2}$	$\frac{1}{2}$	1.00
2	1	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{3}{8}$	1.56
2	0	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	1.50
3	1	0	$\frac{1}{2}$	$\frac{1}{2}$	1.00
3	0	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{3}{8}$	1.56
4	0	0	$\frac{1}{2}$	$\frac{1}{2}$	1.00

the suspect coins is counterfeit—otherwise, the one that weighed more). If experiment $\alpha_2^{(3,0)}$ has outcome R , then one of the three coins lying on the right beam is counterfeit and is heavier than the genuine ones. Comparing the weights of two of these coins (by a third weighing), we can find the outcome of β (either the heavier one is counterfeit, or if they are equal, the third coin is counterfeit). Similarly, we can also analyze the cases in which experiment $\alpha_2^{(2,1)}$ or $\alpha_2^{(1,2)}$ has outcome L .

Case II. *One beam of the balance (say, the right one) is heavier for the first weighing.* In this case, either one of the four coins on the right beam is counterfeit and heavier than the others, or one of the four coins on left beam is counterfeit and lighter. In the second weighing, we can place on the right beam i_1 coins from the right beam and i_2 coins from the left beam, and on the left beam j_1 coins from the right beam, j_2 from the left beam and $(i_1 + i_2) - (j_1 + j_2)$ definitely genuine coins not on the balance during the first weighing (experiment $\alpha_2^{(i_1, i_2; j_1, j_2)}$; assume that $i_1 + i_2 \geq j_1 + j_2$). Here also a table of the entropies of experiments $\alpha_2^{(i_1, i_2; j_1, j_2)}$ can be composed for all possible values i_1, i_2, j_1 and j_2 ; however, since the number of possible variants is fairly large here, it is expedient that some of them be excluded from the very start.

We note that, since the information we expect to gain from the third weighing (experiment α_3) about the outcome of β does not exceed $\log 3$ (because $H(\alpha_3) \leq \log 3$), after two weighings we must be left with *at most* three possible outcomes of experiment β ; otherwise, experiment α_3 will not allow us to determine uniquely the outcome of β . Hence, it is necessary in the first place that the number of suspect coins not on the balance in the second weighing does not exceed 3 (since

in the case of outcome E of experiment α_2 , it is precisely these coins that remain suspect). Thus, we have

$$8 - (i_1 + i_2 + j_1 + j_2) \leq 3, \text{ i.e., } i_1 + i_2 + j_1 + j_2 \geq 5,$$

or, since $i_1 + i_2 \geq j_1 + j_2$,

$$i_1 + i_2 \geq 3, \quad j_1 + j_2 \geq 5 - (i_1 + i_2).$$

Furthermore, if experiment $\alpha_2^{(i_1, i_2; j_1, j_2)}$ has outcome R , then either one of the i_1 'right' coins on the right beam is counterfeit and heavier, or one of the j_2 'left' coins on the left beam is counterfeit and lighter. In exactly the same way, in the case of outcome L , one may suspect that the counterfeit coin is one of the i_2 'left' coins on the right beam, or one of the j_1 'right' coins on the left beam. Hence, we also obtain the following two inequalities

$$i_1 + j_2 \leq 3 \quad \text{and} \quad i_2 + j_1 \leq 3,$$

which must, of course, be satisfied. Finally, it is clear that the inequalities

$$i_1 + j_1 \leq 4, \quad i_2 + j_2 \leq 4 \quad \text{and} \quad (i_1 + i_2) - (j_1 + j_2) \leq 4$$

must also be satisfied.

We now list in the accompanying table all cases satisfying our conditions.

i_1	i_2	j_1	j_2	$p(E)$	$p(R)$	$p(L)$	$H(\alpha_2^{(i_1, i_2; j_1, j_2)})$
2	1	2	1	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{3}{8}$	1.56
2	1	2	0	$\frac{3}{8}$	$\frac{1}{4}$	$\frac{3}{8}$	1.56
2	1	1	1	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{4}$	1.56
1	2	1	2	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{3}{8}$	1.56
1	2	0	2	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{4}$	1.56
1	2	1	1	$\frac{3}{8}$	$\frac{1}{4}$	$\frac{3}{8}$	1.56
3	1	1	0	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{4}$	1.56
1	3	0	1	$\frac{3}{8}$	$\frac{1}{4}$	$\frac{3}{8}$	1.56
2	2	1	1	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{3}{8}$	1.56
2	2	1	0	$\frac{3}{8}$	$\frac{1}{4}$	$\frac{3}{8}$	1.56
2	2	0	1	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{4}$	1.56
3	2	1	0	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{3}{8}$	1.56
2	3	0	1	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{3}{8}$	1.56

Thus we see that here we have not two as in the preceding case but as many as 13 variants of experiment α_2 which contains one and the same maximum information about experiment β (it is perfectly clear that here the information $I(\alpha_2, \beta)$ equals the entropy $H(\alpha_2)$). For any choice of experiment α_2 , this information is found adequate to allow us to determine completely the outcome of β with the aid of one more, that is, the third weighing. Thus, for instance, in the case of outcome E of the experiment $\alpha_2^{(2,1;2,1)}$ one of the two left coins not on the balance in the second weighing is counterfeit. Moreover, we also know that this coin is lighter than the genuine ones; hence, to determine which coin is counterfeit, it suffices to compare the weights of these two coins (or compare one of them with a definitely genuine coin). In the case of outcome R of the same experiment either one of the two 'right' coins on the right beam is counterfeit and heavier or the only 'left' coin on the left beam is counterfeit and lighter. Hence, it is sufficient to compare the weights of the two suspect 'right' coins. The case in which experiment $\alpha_2^{(2,1;2,1)}$ has outcome L can be analyzed in exactly the same manner.

This completes the case of 12 coins. We may now recall the case of 13 coins and show that four weighings are sufficient in this case (we have shown earlier that three weighings cannot suffice here). We place four coins on each beam so that five coins are not on the balance. If one of the two beams is heavier, we have the same situation as encountered by us while analyzing the case of outcome R of the first weighing in the 12 coin problem (with the immaterial difference that we now have not four but five definitely genuine coins). Hence, in this case, three weighings are sufficient to find the counterfeit coin and ascertain whether it is lighter or heavier than the others. If, however, the beams are balanced, then we have to single out the counterfeit coin from not four but from five suspects. Here we may begin by comparing the weight of any one of the suspect coins with that of a definitely genuine coin: if their weights are different, then our problem is immediately solved, otherwise, we are back to the case of four suspect coins, and then, with the aid of two weighings, we can determine the counterfeit coin and ascertain it to be lighter or heavier than the others (see Case I on p. 112 and onwards).

The next problem now generalizes the conditions of Problem 28.

Problem 29. *There are n coins of the same denomination, of which one is counterfeit and is either lighter or heavier than the rest. What is the least number k of weighings on a beam balance that is necessary to find the counterfeit coin and ascertain whether it is lighter or heavier than the others (see [62], Problem 6b).*

This problem is related to the examination of experiment β which may have $2n$ outcomes. It is natural to consider all these outcomes to be equally probable; hence, the entropy $H(\beta)$ equals $\log 2n$. Moreover, the entropy of the

experiment $A_k = \alpha_1 \alpha_2 \dots \alpha_k$, consisting of successive k weighings does not exceed $k \log 3 = \log 3^k$; hence, we must have

$$2n \leq 3^k, \quad \text{that is,} \quad n \leq \frac{3^k}{2},$$

or since n and k are positive integers and 3^k is odd,

$$n \leq \frac{3^k - 1}{2}.$$

In other words,

$$k \geq \log_3(2n + 1) = \frac{\log(2n + 1)}{\log 3}.$$

Thus, say, if $n > (3^3 - 1)/2 = 13$, then the counterfeit coin *cannot* be found with less than three weighings.

It is also easy to see that, even in the case in which $n = (3^k - 1)/2$, k weighings *do not always enable* us to find the counterfeit coin and ascertain whether it is lighter or heavier than the others. (For example, when $n = 13$, the counterfeit coin may not be found in all cases from three weighings.) The proof of this is quite similar to the one given above for the particular case $n = 13$ and $k = 3$ (see the start of the solution of Problem 28). Indeed, for evaluating the entropy of experiment $A_k = \alpha_1 \alpha_2 \dots \alpha_k$ we have so far proceeded from the fact that the entropy of each individual weighing can equal $\log 3$; in the present case, however, because $n = (3^k - 1)/2$ is not divisible by 3, even the entropy of the first weighing (experiment α_1) cannot attain this value (since the three outcomes of the first weighing can in no way be equally probable). Since $n - 1 = [3(3^{k-1} - 1)]/2$ is divisible by 3, it is clear that in the first weighing it is most advantageous to place the group

$$\frac{n - 1}{3} = \frac{3^{k-1} - 1}{2}$$

of coins on each beam, leaving the remaining group

$$\frac{n + 2}{3} = \frac{3^{k-1} + 1}{2}$$

of coins not on the balance : in this case the probabilities of the three outcomes of experiment α_1 (equal to $(n - 1)/3 : n = \frac{1}{3} - (1/3n)$, $(n - 1)/3 : n = \frac{1}{3} - (1/3n)$ and $(n + 2)/3 : n = \frac{1}{3} + (2/3n)$) are closest to each other and, hence, the entropy $H(\alpha_1)$ of the corresponding experiment is greater than that in any other case.

However, it is plain that the amount of uncertainty that remains after this is such that it cannot be eliminated completely from $k - 1$ weighings. The simplest way to demonstrate this is as follows: we assume that in the first weighing the beams balance. In such a case, one of the group $(n + 2)/3 = (3^{k-1} + 1)/2$ of coins not on the balance is counterfeit, so that we are still left with $3^{k-1} + 1$ equally probable outcomes of β (any of the $(3^{k-1} + 1)/2$ coins not on the balance may turn out to be counterfeit and this coin may be either lighter or heavier than the genuine ones). After ascertaining which of these possibilities eventually occurs, we obtain the amount of information equal to $\log(3^{k-1} + 1)$. This amount of information exceeds the maximum information $\log 3^{k-1} = (k - 1) \log 3$ which can be obtained by $k - 1$ weighings. Similarly, we can show that for any other choice of experiment α_1 (the first weighing) this experiment can have an outcome for which the remaining $k - 1$ weighings will be insufficient for a unique determination of the outcome of β .

Thus, we see that if

$$n \geq \frac{3^k - 1}{2},$$

then k weighings may be insufficient. We now show that, if $n < (3^k - 1)/2$ (i.e., if $n \leq (3^k - 3)/2$; in other words, if $k \geq \log_3(2n + 3) = \log(2n + 3)/\log 3$), then k weighings do suffice.† This conclusion completes the solution of our problem.

We begin with the following auxiliary problem: suppose that in addition to n coins, of which one is counterfeit, we have at least one *definitely genuine* coin; it is required to find the counterfeit coin and ascertain whether it is lighter or heavier than the rest. In this case, as before, we can state that if $n > (3^k - 1)/2$, then k weighings are insufficient (because obviously the amount of uncertainty of the initial experiment does not change because of the addition of a genuine coin). However, we cannot now be certain that, even when $n = (3^k - 1)/2$, the k weighings must be insufficient. In fact, by taking into account the additional genuine coin, we can attain a greater closeness than before among the probabilities of the three possible outcomes of the first weighing and, consequently, gain from this weighing a greater amount of information. For this purpose, it is necessary only to place on each beam a group of $(n + 2)/3 = (3^{k-1} + 1)/2$ coins (one of the $3^{k-1} + 1$ coins used is the additional genuine coin), leaving the remaining $(n - 1)/3 = (3^{k-1} - 1)/2$ suspect coins not on the balance. In this case, it is easy to see that the probabilities of the individual

†This statement has two obvious exceptions: if $n = 1$, then it is impossible to ascertain whether the counterfeit coin is lighter or heavier than the genuine ones (of which in this case there are none); if $n = 2$, then it is impossible to find the counterfeit coin.

outcomes of the first weighing are given by

$$\left[\frac{n+2}{3} + \left(\frac{n+2}{3} - 1 \right) \right] : 2n = \frac{1}{3} + \frac{1}{6n}, \quad \frac{1}{3} + \frac{1}{6n},$$

and

$$\frac{n-1}{3} : n = \frac{1}{3} - \frac{1}{3n},$$

i.e., they are indeed slightly closer to each other than before; consequently, the entropy $H(\alpha_1)$ of experiment α_1 is also slightly larger here. This none-too-large difference, however, suffices to assure the possibility that from k weighings we can find the counterfeit coin and ascertain whether it is lighter or heavier than the others.

For proof of the fact that with even one a fortiori genuine coin at our disposal when $n \leq (3^k - 1)/2$ we can get along with k weighings, it is convenient to make use of mathematical induction. The statement is completely obvious for $k = 1$ (i.e., for $n = 1$). We now assume this to have already been proved for a certain value k and consider the case when $(3^k - 1)/2 < n \leq (3^{k+1} - 1)/2$. If we prove that $k + 1$ weighings are sufficient in this case, then from this stems the validity of our statement in all cases. Let us put in the first weighing on one beam some number x of our n coins and on the other beam $x - 1$ of the n coins plus the lone a fortiori genuine coin; the number of coins not on the scale is then $n_1 = n - (2x - 1)$. The number x is so chosen that we have

$$2x - 1 \leq 3^k \quad \text{and} \quad n - (2x - 1) \leq \frac{3^k - 1}{2},$$

i.e.,

$$3^k \geq 2x - 1 \geq n - \frac{3^k - 1}{2};$$

it is clear that this can be accomplished when $n \leq (3^{k+1} - 1)/2$ (because $n - [(3^k - 1)/2] \leq [(3^{k+1} - 1)/2] - [(3^k - 1)/2] = 3^k$). If in the first weighing, the beams balance, then what remains for us is only to find the counterfeit coin from the number $n_1 \leq (3^k - 1)/2$ of the coins not on the balance. Since we have at our disposal definitely genuine coins too, hence (by the inductive assumption) this can be accomplished with k weighings. If, however, one of the beams is heavier in the first weighing, then we are left with $2x - 1 \leq 3^k$ suspect coins. Moreover, in this case we know that if one out of some a coins is counterfeit, then it is lighter than others, but heavier if it is one of the remaining b coins, where $a + b \leq 3^k$ (if the first beam is heavier, then $a = x - 1$, $b = x$; if the second beam is heavier than $a = x$, $b = x - 1$). In this case also

from k successive weighings we can always find the counterfeit coin (see pp. 110-111).

We now return to our initial $n \leq (3^k - 3)/2$ coins, of which one is counterfeit. In the first weighing we put a group of $(3^{k-1} - 1)/2$ coins on each beam; then, the number of coins not on the balance is†

$$n_1 = n - 2 \frac{3^{k-1} - 1}{2} \leq \frac{3^k - 3}{2} - (3^{k-1} - 1) = \frac{3^{k-1} - 1}{2}.$$

If the beams are equal, then the suspect coins are the $n_1 \leq (3^{k-1} - 1)/2$ not on the balance. Since, in addition, we also have $3^{k-1} - 1$ a fortiori genuine coins, hence, by what has been proved above, from successive $k - 1$ weighings we can find the counterfeit coin and ascertain whether it is lighter or heavier than the genuine ones. If, however, either of the beams is heavier, then the suspect coins are the $3^{k-1} - 1 < 3^{k-1}$ on the balance and we also know that the counterfeit coin is lighter than the others if it is one of the group of $a = (3^{k-1} - 1)/2$ coins but heavier if it is one of the group of $b = [(3^{k-1} - 1)/2] (= a)$ coins. By what has been stated on p. 110, here too we can find the counterfeit coin from $k - 1$ successive weighings. This completes the proof of the assertion made earlier regarding the number of weighings that are necessary.

We further note that for large n the number k , defined by the inequalities

$$k - 1 < \frac{\log(2n + 3)}{\log 3} \leq k,$$

can be quite accurately replaced by the ratio $\log 2n / \log 3$ (in the sense that the ratio $k : \log 2n / \log 3$ rapidly tends to unity for increasing n).

There are, of course, a great variety of different problems related to the determination of counterfeit coins by means of weighings on beam balances. So far, we have considered throughout that only *one* of the coins at our disposal is counterfeit (has a weight different from that of the rest of the coins); however, it may also be assumed, for example, that among the given coins, there are two or more that are counterfeit. Still more difficult are the problems in which the very number of counterfeit coins is also assumed to be unknown.†† We can even consider that the counterfeit coins have two or more different weights. An idea of the new problems arising in this case is given by the next problem due to H. Steinhaus, the Polish mathematician (see [63], p. 42).

†In the case in which n equals $(3^k - 3)/2$, the information $I(\alpha_1, \beta)$ about β contained in our experiment α_1 (first weighing) is exactly $\log 3$.

††For the case of two or more counterfeit coins see, for example, [54] (also [56]). The general case is dealt with in [53] and [57], of which the former contains a more detailed discussion of certain distinctive variants of the counterfeit coin problem (with the indications of the possible applications of those problems) and an extensive bibliography.

Problem 30. *There are four objects of different weights and a beam balance on which the weights of any pair of objects can be compared. It is required to indicate a method that enables us to determine the sequence of the weights of these objects by means of at most five weighings. Show that there is no way to guarantee the possibility of ascertaining the sequence of weights of the objects by means of at most four weighings.*

For 10 objects of mutually different weights there is a method for determining the sequence of the weights of the objects by means of at most 24 weighings (find this method). Can this number of weighings be decreased?

A complete solution of this problem (in which it is obvious that the number of objects can, in fact, be arbitrary) is not known so far; some particular results related to this can be found, for example, in [55] and [58]. There is also a series of other problems of similar type (we shall elaborate on this in the next section); as a rule, they are extremely tedious; however, information theory contributes at least a general approach to their investigation.

3.3. Discussion

In Sections 1 and 2 of this chapter, the concepts of entropy and information introduced in Chapter 2 were applied to analyze certain specific problems of the type of 'mathematical recreations.' In what follows we shall see that reasoning of the same kind is also found to be useful in the solutions of a series of sufficiently serious engineering problems. It will be therefore appropriate to discuss here in depth the general idea of all examples considered; as a result, we arrive also at a more general formulation of the problems, which is highly important for the next chapter.

All the examples of Section 1 and 2 were constructed according to a single scheme. In each of them, we were interested in a certain object from a finite set M of similar objects. For example, in Problems 23 and 24, the set M consisted of some *towns* and we had to determine the town in which the observer O was placed; in Problem 25, M consisted of positive *integers* and in Problem 26 of $\binom{100}{2} = 4950$ *pairs of integers*; in Problems 27—29, M consisted of *coins* and the requirement was to find one of them, namely, the counterfeit coin; finally, in Problem 30, M consisted of all possible *ordered collections of objects* we had at our disposal (such that in the case of 4 objects, M contained all $4! = 24$ possible orderings of these objects) and the problem posed was to find out to which of these arrangements corresponded the weighing sequences of objects, starting from the heaviest and ending with the lightest of them. Using the terminology to which we have become accustomed in the first two chapters of this book, we can assert that we studied the *experiment* β which can have n different outcomes B_1, B_2, \dots, B_n ; also, we denote by M the set of all these outcomes. For separat-

ing the object of our interest (the outcome of β), we make use of *auxiliary experiments* α , each of which can have $m < n$ possible outcomes (the experiments α were either questions which could have two different answers: 'yes' and 'no', or weighings on beam balances which could have three different outcomes: E , R and L). The outcomes of experiments α separate some subset of the set M of the outcomes of β , which enabled us to reject a number of outcomes B_1, B_2, \dots, B_n as 'false' or 'not occurring.' We were required to indicate the least number of auxiliary experiments α that were necessary to find the correct answer to the question we were interested in (i.e., to ascertain the outcome of experiment β) and to describe the precise manner in which this answer could be found most rapidly.

The construction, consistent with the one described above, holds not only for the recreative problems of Sections 1 and 2 but also for many vital problems. Examples of the latter include the problem of efficient coding of messages, which is the foremost concern of this book (see Chapter 4); the problem of sorting out objects according to some criterion; the problem of searching for a word in a dictionary or a requisite book in a large library; the problem of designing an efficient control programme for some objects, say, for lathes in a factory, and so on. Lately, such a wide range of possible applications has evoked a great interest in the themes of Sections 1 and 2 and led to the creation of an elaborate terminology. The system of experiments α that leads to finding the object of our interest (the outcome of experiment β) is called a *questionnaire* and the experiment α itself a *question*; moreover, the questions may differ with respect to both the number of possible answers[†] and, in a series of cases, the 'cost' that characterizes the expenditure involved in the corresponding experiment α or the efforts that have to be put in to obtain a reply (i.e., to find the outcome of α). The problem is to find such a procedure for 'asking questions' (i.e., such a sequence of experiments α) as would lead to the desired answer (the outcome of β) with the aid of the 'shortest' series of questions (in terms of numbers or total 'cost'). There exists an extensive literature devoted to the theory of questionnaires, of which we may just mention [61] by K. Picard, the French mathematician and the Russian review paper [60].

It is clear that in all problems of the sort considered, it is desirable to utilize most expediently the information about the outcome of experiment β which is contained in the results of the auxiliary experiments α . However, it appears that the term 'information' is used here in the commonplace 'everyday' sense

[†]In principle, this does not also exclude the situation in which different possible 'questions' α have different numbers of possible answers; thus, for example, it is possible to conceive a variant of the counterfeit coin problem such that in order to find this coin either questions can be asked of a person who knows which is the counterfeit coin (such an experiment can have *two* answers: 'yes' and 'no'), or weighings of coins can be resorted to (this experiment can have *three* answers: E , R and L).

and not in that more specialized sense which was given to it in Chapter 2. In fact, the quantity I we introduced in Chapter 2 had a purely *statistical* meaning—indeed, its definition was based on the concept of probability. However, in our problems many repetitions of trials do not figure, nor are probabilities involved anywhere; hence, the possibility of applying to these problems the theory developed in Chapter 2 may seem odd at first glance.

The circumvention of the difficulty indicated, which we actually used all the time, consists of the following. Suppose that we solve one and the same problem *many times* (i.e., many times seek the correct answer to one and the same question), where the correct answers are found to be different in different cases and each of the answers has a definite probability of being correct; the corresponding probabilities $p(B_1)$, $p(B_2)$, \dots , $p(B_n)$ are considered arbitrary, but assigned beforehand. In such a case, we can speak of ‘experiment β which consists of finding the correct answer’, the term ‘experiment’ being used here in exactly the same sense in which it was used in the preceding chapter. Experiment β corresponds to the probability table

Outcomes of experiment	B_1	B_2	\dots	B_n
Probabilities	$p(B_1)$	$p(B_2)$	\dots	$p(B_n)$

and its entropy is equal to $-p(B_1) \log p(B_1) - p(B_2) \log p(B_2) - \dots - p(B_n) \log p(B_n)$ which we denote as usual by $H(\beta)$. Since our auxiliary experiments α are always ‘directed straight’ to find the outcome of β in the sense that the knowledge of this outcome completely determines the outcome of α also, so the assignment of probabilities to n outcomes of experiment β enables us to determine also the probabilities of m outcomes of any such experiment α_1 . Hence with reference to α_1 also the term ‘experiment’ can be used in the same sense as in Chapter 2. Furthermore, from the fact that the outcome of β completely determines the outcome of α_1 , it follows that the conditional entropy $H_{\beta}(\alpha_1)$ is zero, and the conditional entropy $H_{\alpha_1}(\beta)$ equals the difference $H(\beta) - H(\alpha_1)$ of the entropies of experiments β and α_1 (see p. 66). But the conditional entropy $H_{\alpha_1}(\beta)$ is the *mean value* of the entropies $H_{A_1}(\beta)$, \dots , $H_{A_m}(\beta)$ of experiment β , corresponding to distinct possible outcomes A_1, \dots, A_m of experiment α_1 . Hence, of these m outcomes, *at least for one outcome* A_i the entropy $H_{A_i}(\beta)$ is found to be not less than $H(\beta) - H(\alpha_1)$; thus, cases are certainly possible for which, after determining the result of trial α_1 , the remaining entropy (the amount of uncertainty) of experiment β is not less than the difference $H(\beta) - H(\alpha_1)$.

It is clear how we can generalize this last reasoning. We arbitrarily choose a sequence of auxiliary experiments (trials) $\alpha_1, \alpha_2, \dots, \alpha_k$, i.e., we consider a certain *compound experiment* $A_k = \alpha_1 \alpha_2 \dots \alpha_k$. We assume also that the in-

dividual experiments $\alpha_1, \alpha_2, \dots, \alpha_k$ are not necessarily independent, i.e., that the results of a preceding trial can influence the conditions for carrying out of succeeding trials; it is even possible that for certain especial outcomes of the first few experiments α , all succeeding experiments become redundant, i.e., can be understood as experiments having a unique fully defined outcome (this means that the compound experiment A_k consists of *not more than* k experiments α_i but not necessarily of exactly k such experiments). In the examples considered, knowledge of the outcome of β always determined the outcome of the compound experiment A_k , so that by using the probabilities of the individual outcomes of β one can also find the probabilities of various outcomes of experiment A_k ; hence here, too, the application of the term 'experiment' to A_k should not cause any confusion.

We also note that if each experiment $\alpha_1, \alpha_2, \dots, \alpha_k$ can have not more than m outcomes, then the total number of distinct outcomes of A_k does not exceed m^k . From the fact that the outcome of β determines the outcome of A_k it follows that the average conditional entropy $H_{A_k}(\beta)$ of β , given the occurrence of the compound experiment A_k , is equal to the difference $H(\beta) - H(A_k)$ of the entropies of β and A_k ; hence for at least one outcome of A_k (i.e., for some specified outcome of k trials $\alpha_1, \alpha_2, \dots, \alpha_k$) the 'residual entropy' of β is not less than $H(\beta) - H(A_k)$.

We now suppose that the difference $H(\beta) - H(A_k)$ is greater than zero. In such a case, for at least one outcome of the compound experiment A_k , there still remains some uncertainty in the outcome of β . In other words, when the entire series of k experiments of α is repeated many times and only those cases are separated out for which all the experiments α had some results specified beforehand, occasionally one or the other answer to our basic question β turns out to be correct. Hence it follows that for the cases in which the compound experiment A_k has the indicated outcome, this outcome does not enable us to determine uniquely precisely which of the answers to the question considered in the problem is correct; therefore, k experiments of α do not suffice here for this purpose.

This very reasoning was used for the solutions of Problems 23—29. In addition, it was also taken into account that an inference on the impossibility of finding the outcomes of β by the k outcomes of experiments α can always be made when *at least for one choice* of the probabilities $p(B_1), p(B_2), \dots, p(B_n)$ of the outcomes of β , the inequality $H(\beta) - H(A_k) > 0$ holds. It is usually found sufficient to consider only the 'most disadvantageous' case for which the entropy of experiment β assumes the maximum value, i.e., for which all outcomes of this experiment are equiprobable

$$p(B_1) = p(B_2) = \dots = p(B_n) = \frac{1}{n};$$

this is precisely what we did in the foregoing when we said that "we shall

consider all the outcomes of β to be equiprobable, since no information is available on these outcomes." It is obvious that subject to such a choice of probabilities for the outcomes of β , the equality $H(\beta) = \log n$ is true. Regarding the compound experiment A_k , an exact calculation of its entropy is often not simple in specific problems; however, in many cases we may succeed by confining ourselves to the simple estimate $H(A_k) \leq \log m^k = k \log m$, which stems from the fact that the number of different outcomes of A_k cannot exceed m^k . In more complicated cases, we evaluate quite accurately the largest 'residual entropy' of experiment β corresponding to the most 'unsuccessful' outcomes of the *first* experiment α_1 , and only after this we use the fact that the entropy of each of the succeeding experiments $\alpha_2, \dots, \alpha_k$ does not exceed $\log m$ (see pp. 111-112 and 116-117). Let us also note that the estimate $H(A_k) \leq k \log m$ leads directly to the important inequality

$$k \geq \frac{\log n}{\log m}. \quad (1)$$

This inequality can, of course, be deduced even without using concepts from information theory; this means that when n different possibilities are involved it is impossible to determine one of them uniquely with the aid of a compound experiment having possibly less than n distinct outcomes.[†] Our foregoing estimate of the necessary number of experiments α frequently reduces to using only the simple inequality (1).

Our basic conclusion on the infeasibility of determining uniquely the outcome of β by the outcome of the compound experiment A_k when $H(\beta) - H(A_k) > 0$ can be justified somewhat differently also. If the outcome of the compound experiment A_k determines completely the outcome of β , then $H_{A_k}(\beta) = 0$ and hence, by virtue of the equality $I(A_k, \beta) = H(\beta) - H_{A_k}(\beta)$, the information $I(A_k, \beta)$ about experiment β contained in experiment A_k must be equal to the amount of uncertainty of β , i.e., $I(A_k, \beta) = H(\beta)$. On the other hand, if the outcome of experiment β also uniquely determines the outcome of the compound experiment A_k , then we have at the same time $I(\beta, A_k) = H(A_k)$. Thus, *if the compound experiment A_k (consisting of not more than k experiments α) enables us in all cases to indicate uniquely the correct answer to a question asked (i.e., to ascertain the outcome of the experiment β), then the equality $H(A_k) = H(\beta)$ must hold.* For instance, in the conditions of Problem 29, it is easy to see that $H(\alpha_1) = \log 3 \approx 1.58$ bits (all outcomes of the first weighing were equally probable); furthermore, for any outcome of the first weighing, the second weighing (experiment α_2) was so chosen that its three outcomes had probabilities $\frac{1}{4}$,

[†]We emphasize that calculating the number of possibilities available here is equivalent to using the simplest uncertainty concept in Hartley's sense (see p. 53).

$\frac{3}{8}$ and $\frac{3}{8}$ and, consequently, $H_{\alpha_1}(\alpha_2) = -\frac{1}{4} \log \frac{1}{4} - \frac{3}{8} \log \frac{3}{8} - \frac{3}{8} \log \frac{3}{8} \approx 1.56$ bits (see pp. 114 and 115); finally, the third weighing (experiment α_3), for the case in which α_2 had an outcome with probability $\frac{1}{4}$, reduced to a comparison of two coins known to have different weights on the beam balance, i.e., had entropy $\log 2 = 1$, but in the remaining $\frac{3}{4}$ of all cases (for either of the two outcomes of α_2 with probability $\frac{3}{8}$) it could have three equally probable outcomes, i.e., had entropy $\log 3$. Hence, we have here $H_{\alpha_1\alpha_2}(\alpha_3) = \frac{1}{4} \log 2 + \frac{3}{4} \log 3 \approx 1.44$ bits and since $H(\beta) = \log 24 \approx 4.58$ bits, we have

$$\begin{aligned} H(A_3) &= H(\alpha_1\alpha_2\alpha_3) = H(\alpha_1) + H_{\alpha_1}(\alpha_2) + H_{\alpha_1\alpha_2}(\alpha_3) \\ &\approx 1.58 + 1.56 + 1.44 = 4.58 \text{ bits} = H(\beta), \end{aligned}$$

as it ought to be. *If, however, the equality $H(A_k) = H(\beta)$ is not satisfied and we have the inequality $H(A_k) < H(\beta)$, then this means that experiment A_k certainly does not allow the correct answer to be indicated uniquely.*

It is also easy to comprehend that the proposition that the *outcome of β completely determines the outcome of the trials α* is not necessary for the last conclusion to be true. If this proposition does not hold, then the assignment of probabilities to individual outcomes of β does not enable us to predetermine uniquely the probabilities of all outcomes of the auxiliary experiments α . Hence, while assuming that the experiments for the determination of the outcome of β by the outcomes of experiments α , are carried out *many times*, it is further necessary here to assign also the probabilities to the latter outcomes (of course, they should be such that their values do not contradict the already assigned values of the probabilities of the outcomes of β). In this case, as before, if the compound experiment $A_k = \alpha_1\alpha_2 \dots \alpha_k$, consisting of not more than k trials of α , completely determines the outcome of β , then the information $I(A_k, \beta) = H(\beta) - H_{A_k}(\beta)$ equals the entropy $H(\beta)$. On the other hand, since we always have $I(A_k, \beta) = H(A_k) - H_{\beta}(A_k) \leq H(A_k)$, the inequality $H(\beta) \leq H(A_k)$ must hold. Thus, as earlier, *if*

$$H(A_k) < H(\beta),$$

then the outcome of the compound experiment $A_k = \alpha_1\alpha_2 \dots \alpha_k$ cannot in all cases uniquely determine the outcome of β . This conclusion enables us to obtain an estimate of the least number k of trials α that permits us to determine the outcome of β . However, in the case under consideration here, the estimate so obtained is found to be usually strikingly less precise than in the case in which the outcome of β uniquely defines the outcomes of all trials of α . This is related to the fact that in the former case the trials of α are not directed straight to finding the outcome of β and, consequently, the information $I(A_k, \beta)$ with respect to β , contained in the k trials $\alpha_1, \alpha_2, \dots, \alpha_k$ is not equal to but less than the entropy $H(A_k)$.

Let us, for example, assume that in the conditions of Problem 29 (see p. 116) we are not required to find out whether the counterfeit coin is heavier or lighter than the genuine ones (we are only required to *indicate* that this is a counterfeit coin). We assume that any one of the n existing coins may, with a definite probability, turn out to be counterfeit. In this case, we can calculate the probabilities of all outcomes of experiment β . If, in addition, we assume that the counterfeit coin has a definite probability of being heavier or lighter than the rest of the coins, then we can determine also the probabilities of all outcomes of any trial α , which permits us to speak, with complete legitimacy, of the entropy of experiments α and β and the information contained in either of them with respect to the other. In particular, if we consider that all outcomes of experiment β are equally probable (i.e., that each of the n coins has the same probability of being counterfeit), then the entropy $H(\beta)$ of experiment β equals $\log n$. On the other hand, the entropy of each of the experiments α does not exceed $\log 3$ (since an experiment of this sort, as before, can have three distinct outcomes: E , L and R) and the entropy of the compound experiment $A_k = \alpha_1 \alpha_2 \dots \alpha_k$ does not exceed $k \log 3$. This implies that the least number k of weighings required to find the counterfeit coin must satisfy the inequality

$$k \geq \frac{\log n}{\log 3}. \quad (2)$$

This estimate leads to the number k being smaller than that in an analogous estimate of the least number of weighings necessary to determine the counterfeit coin and find out *whether it is lighter or heavier than the rest*; the inequality then has the form

$$k \geq \frac{\log 2n}{\log 3} \quad (3)$$

(because here experiment β has $2n$ distinct outcomes, since each coin may turn out to be either lighter or heavier than the rest). However, estimate (3) is rather exact: thus, for $k = 3$, estimate (3) gives $n \leq 13$ and, as we know, the maximum number of coins from which (in three weighings) we can separate a counterfeit coin and determine whether it is lighter or heavier than the rest is 12 (see Problem 28). In contrast to this, estimate (2) is highly inaccurate: it yields only the inequality $n \leq 27$ for $k = 3$, whereas we can actually verify that the maximum number of coins from which the counterfeit coin can be separated in three weighings without ascertaining whether it is lighter or heavier than the rest is only 13. This is explained by the fact that here the experiments α (i.e., the weighings of the coins) are not directed straight to the determination of the outcome of β (they contain 'extraneous' information, namely, information about the weight of the counterfeit coin). Hence the contribution of each such

experiment to the information accumulated about the outcome of β is significantly less than $\log 3$ and, consequently, the number of experiments α has to be considerably *larger* than $\log n/\log 3$.

Let us now revert to the question of how it can be established that the outcome of the experiment β we are interested in *can* indeed be uniquely determined by means of not more than k auxiliary experiments α . We have spoken so far only of the *proofs of the infeasibility* of finding the outcome of β with the aid of sufficiently small number of trials α . Similar 'proof of feasibility' involves indicating explicitly the most expedient chain $\alpha_1, \alpha_2, \dots, \alpha_k$ of auxiliary experiments, or in other words, indicating the appropriate compound experiment A_k . Of course, the 'solution' obtained in this case does not include the entropy and information concepts. These concepts nevertheless play an important heuristic role, since they are handy in determining most rapidly the appropriate chain of trials. In fact, the objective of our trials is to ascertain the outcome of experiment β , i.e., to obtain complete information about this experiment. Hence, it is natural to select these trials in such a manner that they contain maximum possible information about the outcome of β . A rigorous method of solving the problem is to enumerate all the possible compound experiments $A_k = \alpha_1 \alpha_2 \dots \alpha_k$, evaluate the information $I(A_k, \beta)$ for each of them, and select those A_k which satisfy the equality $I(A_k, \beta) = H(\beta)$. In the case in which the outcome of β uniquely determines the outcome of all trials α , the evaluation of information is considerably facilitated by the fact that here we should have $I(A_k, \beta) = H(A_k)$. However, since it is inconvenient to operate directly with the compound experiments A_k , in practice it is usual to start with the determination of that auxiliary experiment α_1 (1st trial) that contains the greatest amount of information $I(\alpha_1, \beta)$ about the outcome of β , then select the second trial α_2 (depending in general on the outcome of α_1) such that the information $I(\alpha_1 \alpha_2, \beta)$ is the maximum possible, and so on. This is exactly what we did in solving Problems 23–29.†

We have assumed, throughout Sections 1 and 2, that all outcomes of experiment

† We give one instructive example to illustrate the complications that may arise in the realization of this programme for those cases in which $H_\beta(\alpha) \neq 0$ and the trials α are not directed wholly to the determination of the outcome of experiment β . Suppose we must determine by means of weighings on a beam balance whether a single counterfeit coin among four given coins is lighter or heavier than the rest (however, it is not required to find the counterfeit coin). It is obvious that here every weighing α_1 contains *zero* information with respect to the experiment β we are interested in (since in any outcome of experiment α_1 the probability that the counterfeit coin is lighter and that it is heavier than the genuine coin is in no way altered), i.e., any choice of α_1 leads to one and the same result, which is discomforting at first sight. However, the obligatory equality $I(\alpha_1, \beta) = 0$ does not at all mean that the auxiliary experiments α are useless: experiment α_1 supplies directly no information about β , but it then enhances the suitability of the subsequent trials for this purpose. In fact, it is easy to see that after placing one or two coins on each beam (i.e., choosing experiment α_1 quite arbitrarily), we immediately arrive at the position in which by means of one more weighing (experiment α_2) we can uniquely determine the outcome of β .

β are *equally probable*. This assumption means that all outcomes of β are considered to be equivalent. This is completely natural since it is necessary that a larger number of trials are not involved in order that we are able to determine the outcome of β , *no matter what this outcome may be*. Clearly, the route for determining the outcome of β , satisfying this condition, leads in general to the compound experiment A_k consisting of all cases (i.e., for every outcome of β) of roughly one and the same number of individual trials α . Let us recall, for example, Problem 25 of Section 3.1, which required us to ascertain, using a minimum number of questions, which of the numbers from 1 to 10 was thought of by a certain person. In the solution of this problem, we proposed to clarify in the first place whether the unknown number x exceeded 5 (trial α_1); next, depending on the outcome of α_1 , we recommended determining whether or not the number x was greater than 7 or 3 (trial α_2); further, taking account of the outcome of α_2 , it was possible to inquire whether or not x was greater than 8 or 6, or 4 or 1 (trial α_3); finally, if the three trials α_1 , α_2 and α_3 failed to classify the value of x , then we inquired further concerning whether or not x was greater than 9 or 2 (trial α_4). In all cases, it was necessary to make use of not more than four questions to determine the number x . Moreover, if x happened to equal one of the numbers 2, 3, 9 or 10, then the number of questions needed was exactly four and in the remaining six cases it was three. Clearly, had we asked at the very start whether or not the number x was equal to, say, 10, then we would have a definite chance of doing the trick with *exactly one question*; however, in most cases, we have to invest a larger outlay than a series of four questions, which makes such a method of determining the outcome of β less profitable.

We now note that had we started with the question of whether or not the unknown number x exceeded 8, then we would have had a chance of finding x by asking two questions in all (if this number x were 9 or 10), and at the same time we would not have needed in any case to ask more than four questions (because if after the first question we found that the number x did not exceed $2^3 = 8$, then we could have determined it by means of three more questions; see the solution of Problem 25). Thus, a cursory glance suggests that such a method of finding the unknown number x is more profitable than that proposed in Section 3.1. However, this is a rather hasty conclusion. In fact, if we do not consider the length of the *longest* chain of trials as a single criterion, but determine the value of a method for finding x , taking into account also the fact that in some cases this method leads faster to the goal, then even with respect to the method developed above we must consider the fact that in many cases it allows us to find x with the aid of three and not four questions.

In order to compare the 'merits' of both methods used for the solution of Problem 25 in the light of the foregoing new approach to it, we assume that the trials to find the unknown number x are repeated *many times*, and consider as before that the probability of all ten numbers being thought of is the same. In

the first method for the solution of the problem, we have to ask three questions altogether in roughly $\frac{3}{10} = \frac{3}{5}$ of all the cases and four questions in $\frac{4}{10} = \frac{2}{5}$ of the cases (when x equals 2, 3, 9 or 10). Thus, the *mean value* of the number of questions asked here is

$$\frac{3}{5} \times 3 + \frac{2}{5} \times 4 = \frac{17}{5} = 3.4.$$

The second method for the solution of the problem assures the determination of x with the aid of two questions in $\frac{2}{10} = \frac{1}{5}$ of the total number of all trials (when x equals 9 or 10), whereas in the remaining $\frac{8}{10} = \frac{4}{5}$ cases, four questions must be asked. Hence, *mean value* of the number of questions asked here is given by

$$\frac{1}{5} \times 2 + \frac{4}{5} \times 4 = \frac{18}{5} = 3.6.$$

Thus, on the *average* the second method of finding x is slightly less advantageous than the first. This situation has a general character, which can be verbalized as follows: *whatever the number n , there does not exist a method for the solution of Problem 25 which, on the average, would be more advantageous than the one given on pp. 105-106.*

The preceding conclusion lends a new insight into the problems considered in Sections 3.1. and 3.2. It also makes more transparent the idea underlying the use of the concepts of entropy and information for the solution of such problems. It is clear that the application of these concepts, having an essentially statistical character, is completely relevant only to those cases in which the problem to be solved itself has a statistical character, i.e., it is related to the many repetitions of one and the same trial. The whole point is that we can also understand the foregoing Problems 23-29 in exactly this way if we are interested not in the exact number of trials α that are required to somehow determine a single outcome of experiment β but rather in the *mean value* of this number when the indicated experiment is repeated many times. If, in this case, it is further stipulated that all outcomes of β are considered to be equally probable, then for a choice of the trials $\alpha_1, \alpha_2, \dots, \alpha_k$ such that the mean value of their numbers is the least, the number of trials performed is nearly the same for all outcomes of β . Hence, also the largest value of the number of trials involved here is, in general, the least possible.

Let us now try to *do away with the condition wherein the outcomes of β are considered to be equally probable*. By way of an example, we recall Problem 25 but now make its structure slightly more complex. Suppose that someone thinks of a definite number x that can take one of n values. It is required to find this number by asking some 'yes-or-no' questions. It is further assumed that we

have, beforehand, some information about the number x because of which we must consider the n possible values of this number to be not equally probable, i.e., some of them are more likely to turn out to be the number thought of than the others.† How should the questions be asked in this case?

It is clear that, if none of the n values of x is completely excluded by the information available to us (otherwise we should have spoken not of n , but of a smaller number of possible values of x), then the least number of questions, which guarantee *in all cases* the determination of the number x , is defined as before by the inequalities (1) of Section 3.1 (p. 106), and it is necessary here to ask questions exactly in the same way as stated previously. Indeed, if there were a sequence comprising fewer questions, that would enable us in all cases (i.e., independently of the answers to these questions) to determine uniquely the number x , then this would contradict the result of Problem 25. However, this still does not imply that it is always expedient to act in exactly the same way as in the case in which all values of x are considered to be equally probable; after what has been stated above this ought to be perfectly clear. Thus, for instance, if there is a very large probability that the number thought of has some definite value x_0 (say, if this probability is 0.99 or still higher), then it is obviously reasonable to ask in the first place whether or not x is equal to this number x_0 , in spite of the fact that in the case of a negative answer we waste one question without deriving much profit (the set of possible values of x is simply decreased by one). Moreover, in the general case it is profitable that *every time the set of possible values of x is partitioned into two such parts that the 'probabilities' of the 'number thought of' to belong to either of these parts be as nearly equally likely as possible*. This partition ensures that the entropy of experiment α (when α consists of asking whether or not x belongs to one of these parts) will be the largest possible and, consequently, also ensures that the information contained in α with respect to the experiment β we are interested in will be the maximum possible. It is true that here we are not yet able to guarantee the minimality of the *number* of questions that we may require in the most unfavourable case but, on the other hand, here the *mean value* of the total number of questions is, in general, less (or, in any case, not greater) than that in any other formulation of questions.

In place of a rigorous proof of the preceding statement we shall confine ourselves here to the verification of it for a simple particular example (see the text in small print at the end of this section). For the most general case, it is comparatively easy to establish only that the mean value l of the number of questions required to determine x is always *not less* than $H(\beta)$ (where $H(\beta)$ is as usual the

†Specifically, we may suppose that the number thought of has been written and the person who is to guess has glanced at what was written but is not sure of what he has seen. However, the rigorous sense of this condition is obviously connected with the assumption that, in the process of repeating the procedure of guessing *many times*, some numbers are found to be guessed more often than the others.

entropy of our experiment β).† This result is an extension of the inequality $k \geq \log n$, which is related to the case in which all possible values of x are equally probable; it can be justified by reasoning that is closely analogous to that which led us to the stated inequality. In fact, the information supplied by an answer to a question obviously cannot exceed one bit in any case. Hence, by asking k questions, we obtain information not exceeding k bits. If we now determine many times the number thought of (say, 10,000 times) by asking questions according to some method chosen by us, and if the probabilities that the unknown number coincides with any of the n numbers have assigned values, then the *mean information* given to us in one determination of the number x equals $H(\beta)$ and the total information obtained after 10,000 repetitions of guessing is close to $10,000 H(\beta)$. Of course, the number of required questions may vary here substantially from one determination of x to the other depending on precisely what number x is thought of (it suffices to recall the case in which there exists a definite number x_0 such that the probability that it will be guessed is very large). However, by the very definition, the mean number l of the total number of questions asked in all the 10,000 experiments for finding x is close to $10,000 l$ (this means that, *on the average*, one inquiry about x involves exactly l questions). Hence, we may infer that the inequality

$$10,000 H(\beta) < 10,000 l,$$

i.e.,

$$l \geq H(\beta) \tag{4'}$$

must be satisfied. This is what we are required to prove. Since inequality (4') is vitally important in information theory (see Chapter 4, Section 2 in this regard), we shall deduce, in the sequel, a totally different and highly elegant proof of it, which though more formal, is simpler in concept (see the concluding portion of this section).

All that has been stated above with respect to Problem 25 can easily be carried over also to Problem 27 (pp. 108-109). Here we must generalize the conditions of the problem slightly by considering that different coins have different probabilities of being counterfeit (this can be understood, for example, in the sense that the outward appearance of some coins creates suspicion to a varying extent). In this case, it is most expedient that in each weighing the suspected coins be divided into three groups such that the *probabilities* of the counterfeit coin being found in the two numerically equal groups of coins placed on the right and left beams of the balance and the third group not on the beam balance

†For the case in which n is quite large, and the probability of each individual value of x is small, we can also show that this *mean value* is very close to $H(\beta)$ (see Chapter 4).

always remain *as close as possible to each other*. In this setting the total number of weighings needed for finding the counterfeit coin may, in an unsuccessful case, be found to be even greater than that given by inequality (2) of Section 3.2 (p. 109); however, the *mean value* of the required number of weighings in this case remains the least. We can also show that this *mean value* l is always not less than $H(\beta)/\log 3$, where $H(\beta)$ is the entropy of the experiment that consists of finding the counterfeit coin, i.e.,

$$l > \frac{H(\beta)}{\log 3} \quad (4'')$$

(see, in particular, the concluding part of this section). Moreover, when the number of coins is large and the probabilities of any one of them being counterfeit are small, the *mean value* l is always very close to $H(\beta)/\log 3$.

We now give a simple example to illustrate the fact that for finding a thought of number x (not exceeding some n) it is of greatest advantage to partition the set of n possible values of x each time into two parts such that the probabilities of x belonging to either of the parts are closest possible to each other.

Suppose that the number n of possible values of x equals 4; in this case the number k defined by inequalities (1) (p. 106) equals 2. Assume now that we are justified in assuming that one value x_0 of x is more probable than the other three values x_1, x_2 and x_3 . Let p be the probability that x equals x_0 and q be the probability that x equals x_i (where i is any of the numbers, 1, 2, 3; $p > q$, $p + 3q = 1$). For the first question, we can ask whether x coincides with either of the numbers x_0 or x_1 ; we could also begin by asking whether x and x_0 are equal. The experiments that consist of asking these two questions we denote by $\alpha_1^{(1)}$ and $\alpha_1^{(2)}$. Since the outcomes of $\alpha_1^{(1)}$ have the probabilities $p + q$ and $2q$, therefore $H(\alpha_1^{(1)}) = -(p + q) \log(p + q) - 2q \log(2q)$. The two outcomes of experiment $\alpha_1^{(2)}$ have, however, probabilities p and $3q$ so that $H(\alpha_1^{(2)}) = -p \log p - 3q \log(3q)$. If $p > \frac{1}{2}$, then obviously the outcomes of experiment $\alpha_1^{(2)}$ have more nearly equal probabilities than the outcomes of experiment $\alpha_1^{(1)}$; if, however, $\frac{1}{2} > p > q$, then we have to compare the differences $(p + q) - 2q = p - q$ and $3q - p$ between the probabilities of the two outcomes of experiments $\alpha_1^{(1)}$ and $\alpha_1^{(2)}$. Since $p - q > 3q - p$ if $p > 2q$, i.e., if $p > \frac{2}{5}$ (because $q = (1 - p)/3$ and $p > \frac{2}{5}(1 - p)$ when $p > \frac{2}{5}$), we infer that when $p > \frac{2}{5}$ we should start with experiment $\alpha_1^{(2)}$ and when $p < \frac{2}{5}$ with experiment $\alpha_1^{(1)}$; when $p = \frac{2}{5}$ it is apparently immaterial which of these two experiments we start with.

If our first question is 'whether x is equal to either of the numbers x_0 and x_1 ', then we partition the set of possible values of x into two *numerically equal* parts; in this case any answer to the first question enables us to find x with the aid of just two questions. If, however, our starting question is 'whether x is equal to x_0 ', then we have a definite chance to find x by a single question; the probability of this being precisely so is equal to the probability that x coincides with x_0 , i.e., equals p . However, if x is not equal to x_0 , then we may not be able to guarantee the possibility of finding x by the succeeding question; the question 'whether x equals the number x_1 ' may be followed by a positive answer (the probability of this being q) but, equally, it may be followed by a negative answer (the probability of this equals the probability that x coincides with x_2 or x_3 , i.e., equals $2q$) and, in the latter case, one more question (the third question) is required. Thus, for the case in which we begin with experiment $\alpha_1^{(2)}$, we have the probabilities p, q and $2q$ of finding x by one, two, and three questions, respectively. Hence

we see that here the *mean value* of the number of questions is given by

$$p \times 1 + q \times 2 + 2q \times 3 = p + 8q.$$

It is easy to verify that $p + 8q < 2$ if $p > \frac{2}{5}$ (because $p + 8q = (8 - 5p)/3$, since $q = (1 - p)/3$). We are thus convinced that it is indeed appropriate to begin with experiment $\alpha_1^{(2)}$ in the case in which $p > \frac{2}{5}$.

Before we conclude this section, we deduce one more rigorous proof of inequalities (4') and (4'') which is not based on any results of Chapter 2 except for the definition of the entropy of an experiment. Instead, we shall make use of the following fact. *Suppose that p_1, p_2, \dots, p_n are any n positive numbers whose sum is 1 and q_1, q_2, \dots, q_n are any other positive numbers whose sum does not exceed 1. Then, we always have the inequality*

$$-p_1 \log p_1 - p_2 \log p_2 - \dots - p_n \log p_n \leq -p_1 \log q_1 - p_2 \log q_2 - \dots - p_n \log q_n. \quad (*)$$

We defer a complete proof of inequality (*) to Appendix I; at present we note only that, for $n = 2$, $p_1 = p_2 = \frac{1}{2}$, $q_1 + q_2 = 1$, this inequality assumes the form

$$-\frac{1}{2} \log q_1 - \frac{1}{2} \log q_2 \geq 1,$$

or differently,

$$\frac{1}{2} \log q_1 + \frac{1}{2} \log q_2 \leq -1 = \log \frac{1}{2}, \quad \text{i.e., } \sqrt{q_1 q_2} \leq \frac{1}{2} = \frac{q_1 + q_2}{2}.$$

Thus, if $p_1 = p_2 = \frac{1}{2}$ and $q_1 + q_2 = 1$, then inequality (*) reduces to the well-known inequality between the arithmetic mean and the geometric mean of two numbers.

Now we recall the experiment β , which has n outcomes B_1, B_2, \dots, B_n giving the probability table

Outcomes of experiment	B_1	B_2	\dots	B_n
Probabilities	p_1	p_2	\dots	p_n

Suppose that, to find which of the outcomes of β actually occurs, a sequence of trials α (the auxiliary experiments), of which each can have m different outcomes, is carried out. We denote, as before, the largest number of trials that may be required for determining the outcome of β by k . Suppose further that n_1, n_2, \dots, n_k are the numbers of those outcomes of β that can be ascertained by means of one (α_1), two (α_1, α_2), \dots , and k ($\alpha_1, \alpha_2, \dots, \alpha_k$) trials. It is obvious that $n_1 + n_2 + \dots + n_k = n$.

We note that the number n_1 of outcomes of β that can be determined with the aid of one trial α_1 obviously does not exceed the number m of outcomes of α_1 :

$$n_1 \leq m.$$

Moreover, $n_1 = m$ only in the case (which is evidently of trivial interest) for which $n = m$ and to each outcome of α_1 there corresponds a unique outcome of β (for example, when in the

conditions of Problem 25 the number of possible values of the number thought of is 2). If however, there exist outcomes of α_1 that do not uniquely determine the outcome of β , i.e., if there are cases in which it is found necessary to carry out the succeeding trial α_2 , then surely $n_1 < m$. In this case, the number of outcomes of α_1 that do not uniquely determine the outcome of β is $m - n_1$. Since the number of outcomes of α_2 equals m , the number n_2 of those outcomes of β that can be determined with the aid of two trials α_1 and α_2 certainly satisfies the inequality

$$n_2 \leq (m - n_1)m = m^2 - n_1m.$$

Quite similarly, if in certain cases we must also carry out the third auxiliary experiment α_3 , then $n_2 < (m - n_1)m$. Moreover, here the experiment α_3 is necessary for not more than $(m - n_1)m - n_2$ outcomes of α_2 . Furthermore, since experiment α_3 itself has m different outcomes in all, it is obvious that

$$n_3 \leq [(m - n_1)m - n_2]m = m^3 - n_1m^2 - n_2m.$$

In exactly the same way we can show that

$$n_4 \leq [(m^3 - n_1m^2 - n_2m) - n_3]m = m^4 - n_1m^3 - n_2m^2 - n_3m,$$

and so on. Finally, for the number n_k of outcomes of β , whose determination involves exactly k trials, it is easy to obtain by induction that

$$\begin{aligned} n_k &\leq [(m^{k-1} - n_1m^{k-2} - n_2m^{k-3} - \dots - n_{k-2}m) - n_{k-1}]m \\ &= m^k - n_1m^{k-1} - n_2m^{k-2} - \dots - n_{k-2}m^2 - n_{k-1}m. \end{aligned}$$

Let us transfer all the terms on the right-hand side here to the left-hand side, except the first term m^k and divide both sides of the resultant inequality by m^k ; then, we obtain

$$\frac{n_k}{m^k} + \frac{n_{k-1}}{m^{k-1}} + \dots + \frac{n_2}{m^2} + \frac{n_1}{m} \leq 1.$$

We denote by l_i (where $i = 1, 2, \dots, n$) the number of trials α that have to be carried out to determine the outcome of β in the case in which this outcome is found to be B_i . In such a case, out of n numbers l_i , there are n_1, n_2, \dots, n_k equal to $1, 2, \dots, k$, respectively. Hence the preceding inequality can also be written in the form

$$\frac{1}{m^{l_1}} + \frac{1}{m^{l_2}} + \dots + \frac{1}{m^{l_n}} \leq 1.$$

We now recall that for inequality (*) to be valid the only requirement is that the sum of all numbers p_i be 1, and that the sum of all numbers q_i ($i = 1, 2, \dots, n$) should not exceed 1. Hence, we can put into this inequality, in particular, p_i equal to the probabilities of the i th outcome B_i of experiment β and q_i equal to $1/m^{l_i}$, so that

$$\begin{aligned}
 -p_1 \log p_1 - p_2 \log p_2 - \dots - p_n \log p_n &\leq -p_1 \log \frac{1}{m^{l_1}} - p_2 \log \frac{1}{m^{l_2}} \\
 &\quad - \dots - p_n \log \frac{1}{m^{l_n}}.
 \end{aligned}$$

The left-hand side of the preceding inequality is obviously the entropy $H(\beta)$ of β . On the right-hand side we now replace $-\log (1/m^{l_i})$ (where $i = 1, 2, \dots, n$) by $l_i \log m$ to obtain

$$H(\beta) \leq [p_1 l_1 + p_2 l_2 + \dots + p_n l_n] \log m.$$

But by the very definition of the *mean value* (see p. 6) the sum $p_1 l_1 + p_2 l_2 + \dots + p_n l_n$ is exactly equal to the *mean value* l of the required number of trials α . We thus obtain the basic inequality

$$l \geq \frac{H(\beta)}{\log m}. \quad (4)$$

This is also the result we wished to prove. When $m = 2$ (for example, for the case in which experiment α is a 'yes' or 'no' question, it carries over to inequality (4') (because $\log 2 = 1$) and when $m = 3$ (for example, for the case in which α is a weighing on a beam balance) it carries over to inequality (4'').

4

Application of Information Theory to the Problem of the Information Transmission Through Communication Channels

4.1. Basic concepts. Efficiency of a code

In order to illustrate the usefulness of the concepts of entropy and information, which were introduced in Chapter 2, we had analyzed in Chapter 3 a series of 'recreative problems' of such types as are usually popular among high school and undergraduate mathematical enthusiasts. In the present chapter we consider some of the simplest, but in their own right sufficiently serious, applications of these very concepts to the important engineering problem of transmitting information through communication channels. These applications are also shown to have much in common with the already considered 'recreative problems' of finding a thought of number by means of asking questions or a counterfeit coin by means of weighing, so that many arguments from the preceding chapter can be carried over directly to the solution of practical problems in communication engineering.

The starting point is a general scheme for transmitting information through a communication channel—for definiteness, say, through a telegraph line. At one end of the line, the transmitter is fed some *message*, consisting of a series of symbols selected, say, from a set of 27 characters of English (e.g., Latin) alphabet (26 letters and also a 'zero character', a space between words), or of 33 characters of modern Russian alphabet (also including a space), or of 10 digits (in the case of numerical information), or of all the characters and digits taken together. For the transmission of this message in the case of an ordinary wire telegraph, we use current, some characteristics of which the telegraphist can change at his discretion. This enables him to set up a definite sequence of *signals* which are discernible by the other telegraphist at the receiving end. The simplest distinguishable signals that are extensively used in practice are those of *switching on the current pulse* (i.e., switching on the current at some well-defined instance) and cutting off the current, thus creating a *pause* (the cut-off of current at the same time). Any message can be transmitted by means of these two signals, if every character or digit is agreed upon to be replaced by a definite combination of current pulses and noncurrent pauses.

In communication engineering, the rule that associates some combination of

signals with each message to be transmitted is usually called a *code* (in the case of a telegraph, for example, a telegraphic code) and the operation of the transmission of message into a sequence of distinct signals, the *coding* of message. A code using only two distinct elementary signals (such as, current pulse and non-current pause) is called a *binary code* and a code using three distinct elementary signals a *ternary code*, and so on. In telegraphy, in particular, a number of distinct codes are used, the most noteworthy of which are *Morse code* (Morse alphabet) and *Baudot code*. In the Morse code, we associate with every letter or digit of the message some sequence of short-duration current pulses (dots) and three times longer current pulses (dashes), separated by short duration pause of the same length as a 'dot.' Moreover, the gap between the letters (or digits) is recorded by a special separation mark, a long pause (of the same length as a 'dash'), and the gap between the words by a pause that is twice as long as that between individual characters. Although this code uses only current pulses and pauses, it can be regarded as a ternary code because every encoded piece of information in this case naturally decomposes into a collection of the following three relatively large 'elementary signals'—dots, to each of which (within the letter or digit encoded) is invariably added a dot-length pause, a dash followed in each case by a short-duration (dot-length) pause, and lengthy pauses (dash-length) that separate the individual characters. Morse code is at present used only when the basic telegraphic channels are damaged, and also in short-wave radio-telegraph, which finds many important applications.

The binary Baudot code is mostly used in the ordinary letter-printing telegraphic apparatus installed at all the modern telegraphic offices. This code associates with every character some sequence of five elementary signals—consisting of current pulses and pauses of the same length. Since all the characters are transmitted here by a combination of signals of the same length (codes having this property are called *uniform codes*), no special mark is necessary in the Baudot code for separating one character from another, because it is known beforehand at the receiving end that after every five elementary signals one character terminates and the succeeding one starts (in the receiving apparatus such partition of the sequence of signals into combinations of five signals is usually carried out automatically). Since by combining the two possibilities of the first signal with the two possibilities of the second, two of the third, two of the fourth, and two of the fifth, we can compose altogether $2^5 = 32$ different combinations, the Baudot code in its simplest form allows us to transmit 32 distinct characters.†

†Since 32 combinations for the transmission of all characters and digits are found to be inadequate, in apparatus using the Baudot code there are two registers; after switching on the registers the same combination is used for the transmission of one more character. The number of possibilities is thus almost doubled, which enables us to transmit all letters, digits and punctuation marks. In the case of a single register such possibilities are, however, admitted in a code that associates with every letter or digit a combination of *six* elementary signals; such codes are also used sometimes in telegraphy.

In certain types of telegraphic apparatus, besides simple on and off currents, it is also possible to reverse the current direction. This affords an opportunity to discard current pulses and pauses and instead use as basic signals the current pulses in two different directions or even use simultaneously three distinct elementary signals of the same length: the current pulse in one direction, the current pulse in the other direction, and a pause. Still more complex types of telegraphic apparatus are also possible, in which the pulses are differentiated not only by the direction but also by the amplitude of the current. This gives us an opportunity to further enlarge the number of distinct elementary signals. An increase in the number of such signals allows us to make a code more compact (i.e., to decrease the number of elementary signals required for the transmission of the given information or to transmit by means of signals of the same length a significantly larger number of different 'characters'). However, at the same time it complicates and makes costlier the transmission system, so that in practice it is always preferable to use a code with a smaller number of elementary signals.

In a radiograph, in place of the current amplitude, some parameters of a radio-wave (sinusoidal oscillations of high frequency) are varied, i.e., the elementary signals here have a different sense. However, in this case also every character to be transmitted is replaced by some sequence of elementary signals that are discernible at the receiving end of the channel. A similar situation holds also in the majority of other communication channels. This is discussed in greater depth in Secs. 4.3 and 4.4.

We now dispense with engineering details and formulate the fundamental mathematical problem we have to deal with in communication engineering. Suppose that there is a message written by means of some 'alphabet' containing 'character' (say, 27 English characters, or 33 Russian characters, or 10 digits, or all the characters and digits, or characters, digits and punctuation marks and so on). It is required to 'encode' this message, i.e., to indicate a rule which would associate with every such message a definite sequence of m different 'elementary signals' that make up the 'alphabet' for transmission. How this can be made most advantageous?

In the first place, we must clarify in what sense the term 'advantageous' is understood here. *We consider that the more advantageous a coding is the fewer are the elementary signals that have to be used for the transmission of message.* If it is assumed that each elementary signal takes up the same time, then the most advantageous code is that which allows us to spend the least time in message transmission. Since the installation and maintenance of communication channels are usually very expensive (and in the case of radio communication, where a slightly different position holds, an indiscriminate increase in the number of communication channels is impossible because this can give rise to interference between adjacent channels), it is surely of great importance to move on to a more advantageous code that enables us to use a given communication channel

more efficiently.

We shall now try to analyze somewhat in detail such sorts of codes as are generally used. For definiteness we shall assume for the time being that $m = 2$ (i.e., our code is binary). In addition, we restrict ourselves only to the case of one-letter coding, i.e., to the case of codes that are suitable for transmitting each individual letter of a message (we shall speak later about the opportunities opened up by the rejection of this last restriction). In such a case the coding obviously is such that each of the n 'letters' of our 'alphabet' is assigned some sequence of the two elementary signals, called the *code word* associated with the corresponding 'letter.' If we choose to ignore the physical nature of the elementary signals to be used, we can replace them by the digits 0 and 1, i.e., consider all code words as some sequence of these two digits. For assigning a code it is necessary to enumerate n such sequences to be associated with n existing 'letters.' Besides, not every n distinct sequence of the digits 0 and 1 is suitable for practical use in a binary code; it is still necessary to assure that the encoded information can be *uniquely decipherable*, i.e., in a long sequence of digits 0 and 1 assigned to a multiletter message, it should always be possible to understand where the code word of one letter ends and that of the succeeding letter starts. It is quite simple to achieve this if, as in Morse code, a special separating symbol is introduced (in the engineering literature, such a symbol is sometimes called a 'comma'), which is distinct from all other code words and easy to distinguish, and this symbol is transmitted between the code words of each two 'letters.' It is, however, plain that this method can hardly be advantageous, since here the number of 'letters' in the message to be transmitted is almost doubled (due to the addition of the $(n + 1)$ th separating 'letters' inserted between every two other letters). Hence in the following we shall be interested only in uniquely decipherable codes without a separating symbol (i.e., 'codes without a comma'). Examples of such codes are, in particular, those codes in which the code words of all letters have the same length (i.e., uniform codes; see the foregoing description of the Baudot code). In addition, there are also many non-uniform codes (containing code words of different lengths) that can be uniquely decipherable and hence do not require a separating symbol. Thus, for example, in the case of a two-letter alphabet (in which $n = 2$) the simplest code without a comma is the uniform code with the code words 0 and 1; however, if we replace the code word 1 by a collection of two digits 11, or 10, or 01 (but, obviously, not by 00), then such a code is all the more easy to decipher uniquely (in all these cases the code words of the second letter are easily identified in any long sequence of code words of both types by the digits 1 appearing in them).

A more general necessary and sufficient condition that separates *uniquely decipherable codes* among all other collections of n sequences of the digits 0 and 1 can be found in [65] (in this connection see also [64], which deals with the general theory of binary nonuniform codes). For our purpose here it is however adequate to remark only that a nonuniform code can surely be uniquely decipher-

able if *no code word is a prefix of any other longer code word* (so that, for example, if '101' is the code word of some letter, then there cannot be a letter having the code word '1', '10', or '10110'). In fact, if this condition is satisfied, then by reading consecutively the coded script of a message and having before us a list of all code words, it is always possible to tell exactly at what place the code word for one letter ends and that for the succeeding one starts (since here the sequence of elementary signals that starts after the termination of a recurrent code word itself forms a code word only if we cut it off strictly at one definite place).† We further note that a uniform code also obviously satisfies the condition set forth above in italics. As a rule, we shall not consider below codes that do not satisfy this condition. Hence from now on, unless we say otherwise, *by a 'code' we shall mean a collection of n code words associated with n characters of an alphabet for which the condition indicated above is satisfied.*

Let us now take up the question of the relationship of binary coding to Problem 25 on finding a thought of number, which does not exceed n , by means of questions that can be answered 'yes' or 'no.' This relationship is most straightforward. In fact, suppose that we have some binary code; it is convenient to consider that the n 'characters' associated with our code words are all possible numbers from 1 to n . Let us further suppose that it is required to find a thought of number, which does not exceed n . Then we may ask in the first place the question "Is the first numeral of the code word of the number thought of equal to 1?" By way of the second question we may ask "Is the second numeral of this code word equal to 1?" and so on. We thus consecutively determine all numerals of the code word of the number thought of: since none of these words is a prefix of the other, as soon as we arrive at the combination of numbers that make up the code word, we can ascertain the number thought of with complete certainty and announce it. Thus, *every binary code for an n -letter alphabet corresponds to some method of finding out one of the n numbers thought of through 'yes' or 'no' evoking questions.* Conversely, *any method of finding a thought of number* allows us to associate with each of the n numbers a sequence of numerals 1 and 0, where the first numeral shows whether, in the case in which a given number is thought of, the first question is answered as 'yes' or 'no'; the second numeral in exactly the same way indicates the answer to the second question, the third

†A code that satisfies the stated condition is often called an *instantaneous* (or *instantaneously decipherable*) code. This term is due to the fact that, in the case of other uniquely decipherable codes, to determine that we have come to the end of a recurrent code word we have to acquaint ourselves sometimes (or even always) with several succeeding elementary signals, too (that is, the decoding is effected with a lag in comparison to the transmission of information). In the foregoing three examples of nonuniform codes, for a bicharacter alphabet with the code words 0 and 11, or 0 and 10, or 0 and 01, the first two are obviously instantaneous codes but the third one is not (in the third case to clarify the meaning of the digit 0 in a long sequence of digits 0 and 1 that forms the encoded message it is necessary to know also the succeeding digit).

numeral to the third question, and so on. Hence, *any method of finding a thought of number leads to a binary code*. The above formulated condition is obviously always satisfied here because from the fact that our method allows us to indicate uniquely the number thought of through answers to the questions asked, it directly follows that none of the code words obtained can emerge as a continuation of another notation. For example, the presence of the sequence '101' among the code words implies that the answers 'yes', 'no' and 'yes' already completely determine the number and eliminates the possibility of the existence of the code word '10110.'

It is thus seen that the possible binary codes for an n -letter alphabet precisely correspond to all possible methods of determining one of the n numbers thought of by means of 'yes' or 'no' answerable questions. It is now not difficult to understand which code is of utmost advantage. We shall for the present measure the *advantage* (or, more aptly, the *efficiency*) of a given binary code in terms of the maximum number of elementary signals (equivalently, the digits 1 and 0) that are required for the transmission (or writing) of a single character: the less the maximum number, the more efficient is the code. A more precise definition of the 'efficiency' of a code is derived from the calculation of the *average number* of elementary signals corresponding to one character; this definition will be considered in the next section.) In such a case, the problem of constructing a more efficient code coincides with the content of Problem 25. According to the solution of this problem, the greatest number k of elementary signals that make up a character cannot be less than $\log n$, i.e., at most it is defined by the inequalities (1) on p. 106. The necessity of the inequality $k \geq \log n$ is implied by the elementary computations of information. In fact, one letter of an n -letter alphabet can contain $\log n$ bits of information (for this, it is necessary only that all 'letters' of the message be independent of each other and each of them can take all values with the same probability). On the other hand, every elementary signal to be transmitted that takes either of the two values (these being, say, either a current pulse or a pause) cannot contain more than 1 bit of information. Hence, for the transmission of one character not less than $\log n$ elementary signals are needed.

For constructing the *most efficient* binary code we can make use of the solution of Problem 25. Namely, we partition our n 'characters' into two groups as close to being numerically equal as possible; for all characters of the first group we take 1 as the first numeral of the code word and 0 as the first numeral for all characters of the second group. Furthermore, each of these two groups is again partitioned into two closest numerically equal groups and we take 1 as the second number of the code word if the corresponding character is contained in the first of these two smaller groups, and the symbol 0 if it is contained in the second of these groups. Then we partition each of the four already existing groups into two still smaller groups that are numerically equal as closely as is possible and, as before, we choose the third symbol of the code word, and so on.

By what has been stated in Chap. 3.1, we thus arrive at a binary code, for which the maximal number of numerals k in one code word is defined by inequalities (1) on p. 106, so that no code can be more efficient than this one.

This obviously does not imply that there is no other code as efficient, i.e., that there can be only one most efficient code. In particular, it is clear that if we estimate the efficiency of a code consisting of the digits 0 and 1 by the *longest* code word, then we may not at all consider nonuniform codes. In fact, if the code is nonuniform, then we may add to the end of any code word, whose length is less than maximum, a certain number of arbitrarily chosen digits (say, only digits 0) to arrive at a uniform code that has the same maximum length of code word as the original nonuniform code. This deduction is vital for applications since uniform codes have an apparent practical advantage; they are considerably simpler to decode and here the decoding can be easily automatized. We note furthermore that there may be several different uniform codes with the minimum possible length of code words. After emphasizing their great practical importance, we describe here just one more method of constructing such codes, which is in essence quite similar to the code described.

This method involves the use of the *binary number system*. Ordinarily, we use the decimal number system, in which every number is presented as the sum of the exponents of the number 10:

$$n = a_k \times 10^k + a_{k-1} \times 10^{k-1} + \dots + a_1 \times 10 + a_0,$$

where $a_k, a_{k-1}, \dots, a_1, a_0$ are the *digits* of a number which can take values from 0 to 9; the number n is denoted here by a sequence of its digits, i.e., as $a_k a_{k-1} \dots a_1 a_0$. In analogy to this, the number n can also be represented as the sum of the exponents of the number 2:

$$n = b_l \times 2^l + b_{l-1} \times 2^{l-1} + b_1 \times 2 + b_0;$$

here the 'digits' $b_l, b_{l-1}, \dots, b_1, b_0$ must be less than 2, i.e., they can take only the values 1 and 0. In a binary number system the number is denoted by a sequence of appropriate 'binary digits'; thus, for example, since

$$6 = 1 \times 2^2 + 1 \times 2^1 + 0 \times 2^0, \quad 9 = 1 \times 2^3 + 0 \times 2^2 + 0 \times 2^1 + 1 \times 2^0,$$

therefore in the binary number system, the numbers 6 and 9 are written as 110 and 1001, respectively. The numbers can obviously be represented also as the sum of the exponents of any other number m ; we thus arrive at an *m-ary number system*, where the 'digits' can take m values 0, 1, 2, $\dots, m-1$ (such a system will be needed by us later on).

A number k of digits in the usual ('decimal') notation of the number n is obviously defined by the inequalities

$$10^{k-1} \leq n < 10^k;$$

thus, a number in the interval between $10^1 = 10$ and $10^2 - 1 = 99$ has two digits, that between $10^2 = 100$ and $10^3 - 1 = 999$ has three digits, and so on. In analogy to this, a number k of 'digits' in the binary notation of number n is defined by the inequalities

$$2^{k-1} < n \leq 2^k.$$

(Hence it follows directly, in particular, that the number 6 has three digits and the number 9 has four digits in the binary number system.) Therefore, if we write the first n integers starting from 0 (i.e., 0, 1, 2, . . . , $n - 1$), then it is found that with $2^{k-1} < n \leq 2^k$ binary notation of all these numbers will contain not more than k symbols and it is exactly k symbols that are at least once surely required by us. If we now add a definite number of zeros to the beginning of our binary notation of all less than ' k -digit' numbers, we arrive at a uniform binary code for an n -letter alphabet with minimum possible length of code words. Thus, when $n = 10$, say, the corresponding code words are the following combinations that represent an expression in the binary number system of all numbers from 0 to 9 and are supplemented, if necessary, by zeros at the beginning up to four symbols: 0000, 0001, 0010, 0011, 0100, 0101, 0110, 0111, 1000, 1001. All code words for any other n also are equally simple to construct by this method; no preliminary partitioning of a collection of n numbers into smaller groups is involved here.†

We have shown that in the case of an n -letter alphabet the length of code words (i.e., the number of elementary signals contained in them) for the most efficient uniform binary code is the smallest integer k satisfying the inequality $k \geq \log n$. We now note that if $\log n$ is not an integer, then code words of such a length can be used, in general, for the transmission of a *greater amount of information* than that really transmitted in the case of coding a message written by means of an n -letter alphabet. Consider, for instance, the case $n = 10$ (let us say the case for the transmission of numerical information). Every digit of the information being transmitted (written in the usual decimal number system) can take one of ten values, i.e., can contain information at most equal to $\log 10 \approx 3\frac{1}{3}$ bits, which is attainable for the case in which all digits of the message are independent of each other and each of them can take all values with the same probability. Every digit of the encoded message (i.e., every elementary signal being transmitted—say a current pulse or pause) can take either of the two values, i.e., can contain information at most equal to 1 bit (abridged from the words *binary unit*). But the use of uniform binary code involves sending four elementary signals for

†It is easy to see that when n is an integral exponent of 2 (say, $n = 8$, $n = 16$, or $n = 32$), the code obtained with the aid of a binary number system identifies exactly with the one given in the solution of Problem 25. (When $n = 10$, the 'binary code' reduces to the solution of Problem 25, if the solution begins with the question "Does the number thought of exceed 8?"; see p. 129.)

the transmission of one-digit information, and $4M$ elementary signals for the transmission of M -digit information. However, by means of $4M$ binary signals we can transmit information equal to $4M$ bits, i.e., information approximately $\frac{2}{3}M$ bit greater than the maximum information which can be just contained in an M -digit number, i.e., equal to M decimal units of information.

This phenomenon is straightforward to explain. The reason is that when $n = 10$ all symbols in an encoded message are *never* mutually independent and take both possible values with the same probability: these conditions can be satisfied only when $n = 2^k$. In the particular, if we use a code constructed with the aid of the expansion of numbers from 0 to 9 into a binary number system, then in the case in which all digits in the original message are encountered with the same frequency, the digit 0 in an encoded message is encountered $\frac{25}{15} = \frac{5}{3}$ times more frequently than the digit 1 (since it is easily verifiable that in the ten code words written on p. 144 the digit 0 is encountered 25 times and the digit 1 only 15 times). However, for a sequence of given numbers of digits 0 and 1 to contain the largest amount of information, it is necessary that all digits of this sequence take both values with the *same* probability (and be mutually independent).

For the transmission of a long numerical message it is, however, also possible to construct a more advantageous binary code. This necessitates only that we give up letter-wise coding (by 'letters' of which our message consists we of course mean the digits 0, 1, . . . , 9) and use instead the so-called *block codes*, in which code words are associated with 'blocks' consisting of a fixed number of sequential 'letters'. We start with the simplest block of two 'letters', i.e., partition our message into sequential *pairs of digits*[†] and convert into a binary number system not every digit individually but each 'two-digit' number obtained under such partitioning. The number of binary symbols required for writing all two-digit numbers (from 00 to 99 inclusive) is equal to the number of questions needed for finding thought of number within the first hundred, i.e., it equals 7 (see Problem 25, p. 106). Thus, such a system of coding involves for two digits of message an outlay of 7 elementary signals (not $2 \times 4 = 8$, as earlier), i.e., for the transmission of a number containing M digits (for the sake of simplicity M is assumed to be even) it is necessary to send $3.5M$ elementary signals, or $0.5M$ signals less than those in the original system of coding. When it is required to transmit many digits (in the case of M being large) the advantage is found to be quite appreciable.

It is even more advantageous to partition the number to be transmitted into blocks of *three* digits and switch over to a binary number system whenever 'three-digit' numbers are obtained in this process. For the transmission of a 'three-digit' number it is obviously necessary to send 10 elementary signals (see p. 106) so that such a method of coding allows us to transmit a number consisting

[†]Such a partition of a message into sequential pairs of digits is obviously equivalent to its conversion into a hundred-ary number system,

of M digits (in case M is a multiple of three) by means of $\frac{10}{3}M = 3\frac{1}{3}M$ elementary signals. The advantage that can be had from taking recourse to the partitioning of a message into still larger blocks and converting each such block individually into a binary system is quite small in practice (in the passage from a block of three digits to a block of four digits the coding efficiency is even decreased: the transmission of four digits, as it is easy to see, involves $14 = 3.5 \times 4$ elementary signals). Moreover, it is interesting to note that by applying the partitioning into sufficiently large blocks we can further 'condense' our code and *make the ratio of the number of elementary signals in the encoded message to the number of digits in the original (usually decimal) number arbitrarily close to the limit value equal to $\log 10 = 3.32193 \dots$* . In fact, by invoking the partitioning into blocks, of N digits say, we arrive at a code in which every N digit of information involves k elementary signals, where k is an integer satisfying the inequalities

$$k - 1 < \log 10^N \leq k,$$

or, equivalently,

$$N \log 10 \leq k < N \log 10 + 1.$$

Hence, it is seen that in such a code the average number k/N of elementary signals, per decimal digit, cannot differ from the quantity $\log 10$ by more than $1/N$; if we choose N sufficiently large, we can make this difference arbitrarily small (see p. 106).

Clearly, in the foregoing reasoning almost nothing is changed if the original message is not numerical but consists of 'letters' of an arbitrary n -letter 'alphabet' (for example, ordinary English letters, or Russian letters, or letters and digits, or letters, digits and punctuation marks, and so on). In this case, it is also reasonable to use the coding of big blocks of N such 'letters'; for such a coding it is necessary only to expand the first n^N numbers into a binary system. This method makes it possible to achieve the result that the *average number of elementary signals necessary for one letter of message is arbitrarily close to the quantity $\log n$* (a simple calculation of the amount of information substantiates that our average number can never be less than this quantity). It is only in the case in which n is an integral power of 2 (2^k say) that such partitioning into big blocks is found unnecessary; a code can then be made optimal by associating each individual letter with some code word, so that a recourse to block coding confers no advantage. In this context we remark that in a certain sense 'block coding' is always less convenient than 'coding by individual letters': in block coding the decoding is naturally found to be more complex and laborious (the longer the code, the more this is so) and, moreover, it is always effected at the expense of the time required to decode (having received the coded message, it is not possible to determine which is the first letter being transmitted until the succeeding $N - 1$ letters are transmitted).

All the arguments we have adduced easily carry over also to the case in which for transmission we make use of not 2 but m elementary signals (the case of an m -ary code). For constructing a most efficient uniform code the only requirement here is that we use not a binary but an m -ary number system. If n is an integral power of m , then the coding can be completely restricted to each letter of message individually; the number of elementary signals required for the transmission of one letter can also be made here to assume the least possible value, namely the value $\log n / \log m$. However, if n is not an integral power of m , then by assigning each letter of the message individually to a code word we have to send $k > \log n / \log m$ elementary signals for every letter; here k is the least integer greater than $\log n / \log m$. In this case we can construct a more efficient code by using N -letter block coding; if we choose N sufficiently large, we can conclude that the *average number of elementary signals required for the transmission of one letter of a message is arbitrarily close to $\log n / \log m$* . In the particular case $m = 3$, the corresponding arguments will be similar to those deduced in Chap. 3.2 for the determination of the number of weighings on beam balances required to find a counterfeit coin (see pp. 108-109). In fact, since each weighing can have three outcomes, the result of a sequence of such weighings can be represented in the form of a sequence of digits, each of which takes one of the three values[†], i.e., in the form of some number described in a *ternary* system.

4.2. Shannon-Fano and Huffman codes. Fundamental coding theorem

The basic results of the preceding section can be stated as follows: *if the number of letters in an 'alphabet' is n , and the number of elementary signals being used is m , then in any coding method the average number of elementary signals required per alphabet letter cannot be less than $\log n / \log m$; however, it can always be made arbitrarily close to this ratio*, if we only associate directly sufficiently long 'blocks' consisting of large number of letters with the individual code words. From the conceptual view point, this result is obviously linked to the simple arguments stated by Hartley in 1928. It is obviously in no way related to any probabilistic considerations (in Sec. 4.1 the term 'probability' has not at all been mentioned) and actually rests only on an elementary calculation of the number of 'distinct N -letter sequences of an n -letter alphabet' and 'distinct sequences of N_1 elementary signals'. Hence the results of Section 4.1 can hardly claim to establish the importance of information theory for the engineering problem of transmitting messages, of which we spoke in the Preface to the present book.

The results of Sec. 4.1 can indeed be considerably improved if we make use of the concept of entropy, which we introduced in Chap. 2, and take note of the statistical properties of actual messages. As a matter of fact, in Sec. 4.1 we

[†]Since this is taken in a ternary system, these values can be denoted by the digits 0, 1, and 2, but alternatively the letters E , R , and L can also be used (see Chap. 3.2).

characterized quite roughly the efficiency of a code as the *greatest number* of elementary signals involved per letter of the message to be coded, and for this we considered only the simplest codes, i.e., uniform codes. If at the end of that section we talked also of the *average number* of signals involved per letter of message, this was connected only with the fact that there the uniform codes were considered for multiletter blocks and the ratio of the number of elementary signals in a code word to the number of letters in the corresponding block (which we called the average number of elementary signals per letter) could not be an integer. But, in practice, we usually have to deal with messages in which the relative frequencies of different letters differ considerably from each other (it suffices to compare, say, the frequencies of the letters *e* and *y* in any English text; we shall elaborate this in Chap. 4.3). Hence the role of key value must be occupied here by the probabilistic *mean* (or, *average*) *value* of the number of elementary signals involved per letter of message which is defined in accordance with the actual statistical laws characterizing the message to be transmitted.

Let us now examine the problem of the coding of messages that obey definite statistical laws. We consider here only the simplest case of messages written by means of some n 'letters', the frequencies of whose occurrence at any point in the message are completely characterized by the probabilities p_1, p_2, \dots, p_n , where, obviously, $p_1 + p_2 + \dots + p_n = 1$. The simplification which we use here is that the probability p_i of the occurrence of the i th letter *at any point* in the message is assumed to be one and the same irrespective of what letters occur at preceding points; in other words, the successive letters of the message are considered to be *independent* of each other. Factually, in actual message this does not happen very often; in particular, in the English language the probability of the occurrence of a letter essentially depends on the preceding letter (see p. 181 below et seq.). However, if we were to take into rigorous account the mutual dependence between letters, it would highly complicate all our further discussions; at the same time, it is natural to think that this should not alter the results deduced below since, if desired, by 'letters' we can straightaway understand multiletter blocks whose dependence on each other is already comparatively weak.†

We shall consider for the present only *binary* codes; an extension of the corresponding results to codes that utilize an arbitrary number m of elementary signals is as usual quite simple. A brief discussion at the end of the section will suffice for this purpose. We start with the simplest case of codes that associate every 'letter' of a message to an individual code word, a sequence of digits 0 and 1. It has already been remarked above that some method of finding a thought of number

†It can indeed be shown that all the results presented below are preserved for a very wide class of cases, in which the successive letters of a message are dependent on each other (see pp. 161-62 below).

x not exceeding n by means of 'yes-or-no' questions can be associated with every binary code of an n -letter alphabet; conversely, any method of determining such a number leads us to a definite binary code. When probabilities p_1, p_2, \dots, p_n are assigned to individual letters, the transmission of a multiletter message corresponds precisely to the situation described on p. 131 et seq.: the optimal code in such a case is associated with the method for finding a number x for which, with the same n probabilities of the values of x , the average value of the number of questions asked for, is found to be the least. This average value can itself be considered also as the average value of the number of binary symbols (digits 0 and 1) in a code word; in other words, it precisely equals the average value of the number of elementary signals per code letter in the transmission of a multiletter message.

It is now possible to apply directly to our problem the results set forth on p. 131 et seq. According to these results, in the first place *the average number of binary elementary signals per letter of the original message in the encoded communication cannot be less than H* , where $H = -p_1 \log p_1 - p_2 \log p_2 - \dots - p_n \log p_n$ is the entropy of the experiment which consists of distinguishing one letter of the text (or briefly, the entropy of one letter). This directly implies that *for any coding method for writing a long message of M letters we require not less than MH binary symbols*. This statement is immediate from the fact that the information contained in a piece of text of M letters is equal to MH in our case (recall that the individual letters are considered to be mutually independent); at the same time the information contained in one elementary signal (binary symbol) cannot exceed one bit in any way (see p. 132; a variant derivation of the same result is given in small print on pp. 134-36).

If the probabilities p_1, p_2, \dots, p_n are not all equal among themselves, then $H < \log n$. Hence, it is natural to think that by taking account of the statistical laws of a message we can construct a code more efficient than the best uniform code which, according to the results of Sec. 4.1, involves not less than $M \log n$ binary symbols for writing a text of M letters. The procedure used to obtain an optimal code is clear from what was stated on pp. 131-32. It is convenient if, to start with, we arrange all existing n letters in one column in order of decreasing probability. Then, all these letters should be divided into two groups of higher and lower probabilities, so that the total probabilities of the letters of the message belonging to either of these groups should be as close as possible to each other; for letters of the first group we use 1 as the first digit of the code word, and for those of the second group the digit 0. Furthermore, each of the two groups obtained should again be divided into two parts with total probabilities as close as possible to each other; we use either 1 or 0 as the second digit of the code word according to whether our letter belongs to the first or the second of these smaller groups. Then, each of the groups containing more than one letter is again divided into two parts of closest possible total probability, and so on: the process is repeated until we arrive at groups each of

which contains only one letter. Such a method of coding message was first suggested in 1948-1949 by R. Fano and C. E. Shannon independently of each other; hence, the corresponding code is usually called the *Shannon-Fano code* (sometimes simply the *Fano code*).† Thus, for example, if our alphabet contains altogether six letters whose probabilities (in decreasing order) are 0.4, 0.2, 0.2, 0.1, 0.05 and 0.05, then in the first step of division of letters into groups we separate only the first letter (first group), leaving all the rest in the second group. Furthermore, the second letter forms the first subgroup of the second group; however, the second subgroup of that group consisting of the remaining four letters is also again successively divided into parts such that every time the first part consists of only one letter (see the accompanying table). Similarly, in the table

TABLE

<i>No. of letters</i>	<i>Probabilities</i>	<i>Partition into subgroups. The Roman digits signify the numbers of groups and subgroups</i>						<i>Code words</i>
1	0.4	}	I					1
2	0.2	}		}	I			01
3	0.2				}	I		001
4	0.1		II	}		}	I	0001
5	0.05			}	II	}	II	00001
6	0.05					}		00000

on the next page we analyze the case of a richer 'alphabet' containing 18 letters with probabilities 0.3, 0.2, 0.1 (2 letters), 0.05, 0.03 (5 letters), 0.02 (2 letters), and 0.01 (6 letters).

The basic principle of the Shannon-Fano coding method is the following: in the choice of each digit of a code word we wish to ensure that the amount of information contained in it be as large as possible, i.e., that independently of all preceding digits this digit may take either the value 0 or 1 with almost equal probability. The number of digits in different code words is obviously found here to be different (in particular, it varies from one to five in the first example and from two to seven in the second example), i.e., the Shannon-Fano code is nonuniform. It is, however, easy to understand that no code word here can be found to be the prefix of other longer word (this is also clear from the fact that such a code actually coincides with the method described on p. 131 et seq. for solving the problem of finding a thought of number; see pp. 141-42). Hence, an encoded message is always uniquely decipherable. It is quite essential that in the Shannon-Fano code we assign shorter code words to higher probability

†To be more exact, this method of coding was in fact proposed by R. Fano alone; C. E. Shannon, however, offered a slightly different method similar to the one described above.

letters than to low probability ones (because in the successive group divisions higher probability letters are then separated more speedily into the individual one element groups; see the examples analyzed above). As a result, although certain code words may also have quite a significant length here, the *average*

Number of letters	Probabilities	Partition into groups										Code words
1	0.3	}	I	}	I	}	I	}	I	}	I	11
2	0.2											10
3	0.1											011
4	0.1											0101
5	0.05											0100
6	0.03											00111
7	0.03											00110
8	0.03											00101
9	0.03											00100
10	0.03											00011
11	0.02	}	II	}	II	}	II	}	II	}	II	000101
12	0.02											000100
13	0.01											000011
14	0.01											0000101
15	0.01											0000100
16	0.01											000001
17	0.01											0000001
18	0.01											0000000

value of the length of such words is nevertheless found to be only slightly greater than the minimal value H admissible for messages in order to preserve the amount of information in coding. Thus, for the six-letter alphabet example considered above, the best uniform code consists of three-digit code words (because $2^2 < 6 < 2^3$), and hence in that case we assign exactly three elementary signals to each letter of the original message; however, in using the Shannon-Fano code the average value of elementary signals per letter of message is given by

$$1 \times 0.4 + 2 \times 0.2 + 3 \times 0.2 + 4 \times 0.1 + 5 \times (0.05 + 0.05) = 2.3.$$

This value is appreciably less than 3 and is not very far away from the corresponding entropy value

$$H = -0.4 \log 0.4 - 2 \times 0.2 \log 0.2 - 0.1 \log 0.1 - 2 \times 0.05 \log 0.05 \approx 2.22.$$

In analogy to this, for the 18-letter alphabet example considered, the best uniform code consists of five-digit code words (since $2^4 < 18 < 2^5$); however, in the case of the Shannon-Fano code there are letters that are coded by as many as seven binary signals but, on the other hand, the average value of elementary signals per letter is given by

$$2 \times 0.5 + 3 \times 0.1 + 4 \times 0.15 + 5 \times 0.15 + 6 \times 0.06 + 7 \times 0.04 = 3.29.$$

The preceding value is appreciably less than 5 and it does not deviate much from the quantity

$$H = -0.3 \log 0.3 - 0.2 \log 0.2 - \dots - 6 \times 0.01 \log 0.01 \approx 3.25.$$

A special advantage from the Shannon-Fano method is derived when it is used for coding the blocks of several letters and not the individual letters of an alphabet. It is true that here it is nevertheless impossible to exceed the limit value H of binary symbols per letter of message (because, for the case in which the individual letters are independent, the entropy of an N -letter block equals NH and, consequently, in any coding method, there can occur on the average not less than NH binary signals per block). However, even in comparatively unfavourable cases block coding enables us to approach this minimal value rather quickly. Consider, for example, the case in which there are only two different letters A and B with probabilities $p(A) = 0.7$ and $p(B) = 0.3$. Then

$$H = -0.7 \log 0.7 - 0.3 \log 0.3 = 0.881 \dots$$

Here the application of the Shannon-Fano method to the original two-letter alphabet is in fact meaningless: it merely leads us to the following simplest uniform code:

<i>Letter</i>	<i>Probabilities</i>	<i>Code words</i>
<i>A</i>	0.7	1
<i>B</i>	0.3	0

This code requires for the transmission of each letter one binary symbol, this being 13.5% more than the minimal attainable value 0.881 binary digits per letter. However, by applying the Shannon-Fano method to the coding of all possible two-letter combinations (whose probabilities are determined by the multiplication rule of probabilities for independent events, see p. 18), we arrive at the following code:

<i>Letter combination</i>	<i>Probabilities</i>	<i>Code words</i>
<i>AA</i>	0.49	1
<i>AB</i>	0.21	01
<i>BA</i>	0.21	001
<i>BB</i>	0.09	000

The average value of the length of code words here is

$$1 \times 0.49 + 2 \times 0.21 + 3 \times 0.30 = 1.81.$$

Hence, in this case, we need on the average $1.81/2 = 0.905$ binary symbols per alphabet letter, which exceeds by only 3% the value 0.881 binary digits/letter. We obtain still finer results by applying the Shannon-Fano method to the coding of three-letter combinations. This leads us to the following code:

<i>Letter combination</i>	<i>Probabilities</i>	<i>Code words</i>
<i>AAA</i>	0.343	11
<i>AAB</i>	0.147	10
<i>ABA</i>	0.147	011
<i>BAA</i>	0.147	010
<i>ABB</i>	0.063	0010
<i>BAB</i>	0.063	0011
<i>BBA</i>	0.063	0001
<i>BBB</i>	0.027	0000

The average code-word length value is here 2.686, i.e., on the average 0.895 binary symbols per letter of text are needed, which is only 1.5% more than the limit value $H \approx 0.881$ binary digits/letter.

When the difference in the probabilities of the letters A and B is still larger, an approximation to the minimal possible value of H binary digits/letter may be somewhat less rapid, but it is nevertheless reasonable. Thus, when $p(A) = 0.89$ and $p(B) = 0.11$, the value of H is $-0.89 \log 0.89 - 0.11 \log 0.11 \approx 0.5$ binary digits/letter, while the uniform code $A \rightarrow 1, B \rightarrow 0$ (equivalent to the application of the Shannon-Fano code to a set of two existing letters) involves an outlay of one binary symbol for each letter and is twice as long. However, it is easy to verify here that the application of the Shannon-Fano code to all possible two-letter combinations leads to a code in which 0.66 binary digits on the average are necessary per letter. The application of this very code to all three-letter blocks allows one to lower the average number of binary digits per letter to 0.55. Finally, the coding by the Shannon-Fano method of all possible four-letter blocks involves on the average an outlay of 0.52 binary digits per letter, i.e., overall only 4% more than the minimal value of 0.50 binary digits per letter.

The *Huffman code* is closely related to the Shannon-Fano code, but it is more advantageous of the two (see [66]). We now proceed to describe this code. The construction of this code rests on a simple transformation of the alphabet in which the message to be transmitted over communication channels is written. This transformation is called the *contraction* of the alphabet. Suppose that we have an alphabet A containing the letters a_1, a_2, \dots, a_n whose probabilities of occurrence in the message are p_1, p_2, \dots, p_n , respectively; moreover, we consider

the letters to have been arranged in order of decreasing probability (or frequency), i.e., we assume that

$$p_1 \geq p_2 \geq p_3 \geq \dots \geq p_{n-1} \geq p_n.$$

We now agree *not to make a distinction between two least probable letters of our alphabet*, i.e., we consider that a_{n-1} and a_n are *one and the same* letter b of a new alphabet A_1 which obviously contains the letters a_1, a_2, \dots, a_{n-2} and b (i.e., either a_{n-1} or a_n), whose probabilities of occurrence in the message are p_1, p_2, \dots, p_{n-2} and $p_{n-1} + p_n$, respectively. The alphabet A_1 is also called *the alphabet obtained from A by contraction* (or *one-fold contraction*).

The term 'one-fold' carries here the following sense. We arrange the letters of the new alphabet A_1 in order of decreasing probability and carry out the contraction of alphabet A_1 . We then arrive at an alphabet A_2 of which it is natural to say that it is obtained from the original alphabet A by *two-fold contraction* (and from A_1 by a simple or one-stage contraction). It is clear that A_2 contains in all $n - 2$ letters. The continuation of this process leads us to increasingly shorter alphabets so that after $(n - 2)$ -fold contraction we arrive at an alphabet A_{n-2} containing *two* letters in all. By way of an example, our earlier mentioned alphabet containing 6 letters with probabilities 0.4, 0.2, 0.2, 0.1, 0.05 and 0.05 is transformed by successive contractions into the accompanying table.

TABLE

No. of letters	Probabilities				
	Original alphabet	Contracted alphabets			
		A	A ₁	A ₂	A ₃
1	0.4	0.4	0.4	0.4	→0.6 0.4
2	0.2	0.2	0.2	→0.4	
3	0.2	0.2	0.2	0.2	
4	0.1	0.1	→0.2]	
5	0.05	→0.1]]	
6	0.05]]]	

We now agree to assign the code words 1 and 0 to the two letters of the last alphabet A_{n-2} . Furthermore, if code words are assigned to all letters of alphabet A_j , then to the letters of the 'preceding' alphabet A_{j-1} (where, obviously, $A_{1-1} = A_0$ is the original alphabet A), which are also the letters of the alphabet A_j , we assign the *same* code words as they had in the alphabet A_j . However, to the letters a' and a'' of alphabet A_j 'coalesced' into a single letter b of alphabet A_{j-1} we assign the words obtained from the code word of letter b with the addition of digits 1 and 0 at the end; see the following table:

TABLE

No. of letters	Probabilities and code words									
	Original alphabet		Contracted alphabets							
	A		A ₁		A ₂		A ₃		A ₄	
1	0.4	0	0.4	0	0.4	0	0.4	0	0.6	1
2	0.2	10	0.2	10	0.2	10	0.4	11	0.4	0
3	0.2	111	0.2	111	0.2	111	0.2	10		
4	0.1	1101	0.1	1101	0.2	110				
5	0.05	11001								
6	0.05	11000								

It is easy to see that the very construction of the *Huffman code* thus obtained implies that it satisfies the general condition enumerated on pp. 140-141: no code word is here the prefix of another lengthier code word. We also note that the coding of a certain alphabet by the Huffman method (likewise by the Shannon-Fano method as well) is not a uniquely defined procedure. Thus, for example, at any stage of the construction of the code we can obviously replace the digit 1 by 0 and vice versa; then, we obtain two different codes (which obviously differ quite insignificantly from each other and have the same length for all code words). But, apart from this in certain cases we can construct also some Huffman codes that are substantially different; thus, for instance, in the example analyzed above a code can also be constructed according to the accompanying table.

TABLE

No. of letters	Probabilities and code words									
	Original alphabet		Contracted alphabets							
	A		A ₁		A ₂		A ₃		A ₄	
1	0.4	11	0.4	11	0.4	11	0.4	0	0.6	1
2	0.2	01	0.2	01	0.2	10	0.4	11	0.4	0
3	0.2	00	0.2	00	0.2	01	0.2	10		
4	0.1	100	0.1	101	0.2	00				
5	0.05	1011	0.1	100						
6	0.05	1010								

The new code obtained here is also a Huffman code; but the code-word lengths are now entirely different. However, note that the *average number* of elementary signals per letter for both the Huffman codes constructed is precisely identical, getting

$$1 \times 0.4 + 2 \times 0.2 + 3 \times 0.2 + 4 \times 0.1 + 5 \times (0.05 + 0.05) = 2.3$$

in the first case, and

$$2 \times (0.4 + 0.2 + 0.2) + 3 \times 0.1 + 4 \times (0.05 + 0.05) = 2.3$$

in the second case.

Furthermore, it is clear that both the Huffman codes considered are highly effective (the average code-word length here is the same as that obtained above in the Shannon-Fano method). It can also be shown that the Huffman code is the *most effective* of all possible codes in the sense that *in any other method of coding the letters of an alphabet the average number of elementary signals per letter cannot be less than that obtained in the Huffman coding method.* (Let us note that this directly implies also that in any two Huffman codes the average code-word length must be precisely the same—indeed, both happen to be optimal.)

The proof of this *optimality property* of Huffman codes is quite simple. We consider again any n -letter alphabet (we denote it by B , say) containing the letters $b_1, b_2, \dots, b_{n-1}, b_n$ with probabilities $q_1, q_2, \dots, q_{n-1}, q_n$, where

$$q_1 \geq q_2 \geq \dots \geq q_{n-1} \geq q_n, \quad (*)$$

and obtain from it by contraction an $(n-1)$ -letter alphabet (alphabet B_1) containing the letters $b_1, b_2, \dots, b_{n-2}, c$, whose probabilities of occurrence are, respectively, $q_1, q_2, \dots, q_{n-2}, q_{n-1} + q_n = q$. Assume now that we have some system of code words for the letters of alphabet B_1 . Then we carry over this code word system also to alphabet B by retaining the words of all letters that appear simultaneously in both alphabets and forming code words for letters b_{n-1} and b_n by adding 1 and 0, respectively, to the end of the code word of letter c . We now must show that *if the code for the alphabet B_1 is optimal, then the code obtained for the alphabet B in this manner is also optimal.*

To prove the italicized statement we suppose that the code obtained for B is *not optimal* and show that in such a case the original code for B_1 *also cannot be optimal*. In fact, we denote by L_1 and L the *average code-word length* of letters (i.e., the average number of elementary signals per letter) for the codes corresponding to B_1 and B , respectively. It is obvious that

$$L = L_1 + q. \quad (**)$$

Indeed, B_1 and B differ only in that the letter c of B_1 with probability q is replaced in B by two letters b_{n-1} and b_n with the same total probability of occurrence $q (= q_{n-1} + q_n)$; however, the code-word lengths corresponding to these alphabets differ only by an increase per unit of the lengths corresponding to the letters b_{n-1} and b_n in comparison to the length corresponding to the letter c of B_1 . Hence, the relation $(**)$ also follows immediately from the definition of the average code-word length.

It has been assumed that the code corresponding to alphabet B is *not optimal*. In other words, there exists an optimal code other than the one under consideration which associates with the letters $b_1, b_2, \dots, b_{n-1}, b_n$ code words of length (in elementary signals) $k_1, k_2, \dots, k_{n-1}, k_n$ such that in it the average code-word length

$$L' = k_1 q_1 + k_2 q_2 + \dots + k_{n-1} q_{n-1} + k_n q_n$$

is less than L . We can also consider that

$$k_1 \leq k_2 \leq \dots \leq k_{n-1} \leq k_n. \quad (***)$$

In fact, if the letters b_i and b_j (where i and j are any two of the numbers $1, 2, \dots, n$) are such that $q_i > q_j$ (which, because of (*) implies the inequality $i < j$) and $k_i > k_j$, then we simply interchange the code words of b_i and b_j , after which the average code-word length of a letter is further decreased; hence if $q_i > q_j$, then necessarily $k_i \leq k_j$. Now, within a group of letters b_u, b_{u+1}, \dots, b_v (where $1 \leq u < v \leq n$) such that $q_u = q_{u+1} = \dots = q_v$, we can always arrange the letters in such an order that $k_u \leq k_{u+1} \leq \dots \leq k_v$.

From inequalities (***) it follows, in particular, that a codeword having the *greatest* length k_n corresponds to the letter b_n . Furthermore, we can be convinced of the existence of such a letter b_l of the alphabet B , whose code word is *obtained from the code word of b_n by replacing the last elementary signal* (either 1 by 0, or 0 by 1). In fact, if such a code word were altogether absent, then we could simply discard the last elementary signal in the code word of b_n without violating the basic condition given atop p. 141 that defines an instantaneous code (recall that we have no letter whose code word is longer than b_n). But this would again decrease the average length of the code word of a letter which contradicts the assumption of the optimality of the code under consideration.

However, from inequalities (***) and the equality $k_l = k_n$ it follows that inevitably $k_l = k_{n-1}$ (but this does not necessarily imply that $l = n - 1$). We now interchange the code words of b_l and b_{n-1} if $l \neq n - 1$ (if $l = n - 1$, then this step in the reasoning becomes superfluous); here the quantity L' obviously remains unaffected. We now pass from the code for B to the code for alphabet B_1 by retaining the code words of all letters b_1, b_2, \dots, b_{n-2} , and assigning to the letter c the code word obtained from the code words of letters b_{n-1} and b_n with the last digit removed (by which alone these two code words differ). It is obvious that the average code-word length L'_1 of a code for the alphabet B_1 obtained in this manner is related to the average word-length L' of a code for B by the following relation similar to (**):

$$L' = L'_1 + q.$$

Hence, the inequality $L' < L$ implies that

$$L'_1 < L_1.$$

But this also shows that the original code for B_1 is *not optimal*.

We have as a matter of fact already completed the proof of the optimality of the Huffman code. It is indeed clear that the code taken by us for the last alphabet A_{n-2} , which assigns to the two letters of this alphabet the code words 1 and 0, is *optimal*: the average code-word length l of a letter corresponding to it can in no way be decreased. But this implies by what has been proved that the code for alphabet A_{n-3} is also optimal, whence, in turn, follows the optimality of the code for A_{n-4} , and so on till the last code (the Huffman code) corresponding to the original alphabet $A_{1-1} = A_0$, i.e., alphabet A .

The degree of proximity between the average number of binary symbols per letter of a message and the value H attained in the examples considered above can be further increased arbitrarily by taking recourse to the coding of increasingly lengthier blocks. This flows from the following general statement which we shall hereafter call the *fundamental coding theorem*†: *in coding a message segmented into N -letter blocks, it is possible by choosing N sufficiently large to assure that the average number of binary elementary signals per letter of the*

†To be more exact, it should be designated as the *fundamental coding theorem for the noiseless channels*. The extension of this result to the problem of the most advantageous coding, taking account of the impact of noise, is considered in Sec. 4.4.

original message is arbitrarily close to H (in other words, arbitrarily close to the ratio of the amount of information H contained in a letter of the message to 1 bit, i.e., to the greatest amount of information that can be contained in one elementary signal). Differently, this can also be formulated thus : *a quite long message of M letters can be encoded by means of the number of elementary signals arbitrarily close to* (but obviously in no case less than) MH , *if only this message is divided beforehand into sufficiently long blocks of N letters and separate code words are straightaway associated with all blocks.* We further note that it is not by accident that we have not stated anything here as to precisely how we should construct N -letter blocks: as seen in the following, the methods for block coding may be highly diverse (thus, for example, it is possible to follow either the Huffman or the Shannon-Fano coding method, but these are by no means the only possibilities open to us). Thus, the partitioning of a message into quite lengthy blocks plays a central role in the construction of an optimal code. It will be seen in Section 4.4 that direct block coding is of considerable advantage in the case of noisy channels, too (though the coding method itself has to be substantially modified in that case).

In view of the crucial importance of the fundamental coding theorem, we shall now give *two* completely different versions of its proof (both due to C. E. Shannon). The first essentially rests on the use of the Shannon-Fano coding method though, as we shall see later, a direct appeal to this method is not made in the proof. It is presumed for the present that under the successive divisions of the collection of letters to be coded (which can also be understood as entire 'blocks') into smaller groups, which forms the basis of the Shannon-Fano coding, we succeed each time in attaining the result that the total probabilities of both the groups obtained are precisely *equal* to each other. In such a case, the first, second, . . . , l th divisions yield the groups whose probabilities sum to $\frac{1}{2}$, $\frac{1}{4}$, . . . , $1/2^l$, respectively. The l -digit code word has here those letters which were found to have been extracted in the one-element group after exactly l divisions, i.e., the letters whose probability is $1/2^l$. In other words, subject to the fulfilment of this condition *the code-word length l_i is related to the probability p_i of the corresponding letter by the formula*

$$p_i = \frac{1}{2^{l_i}}, \quad l_i = \log \frac{1}{p_i} = -\log p_i.$$

In fact, our condition can be precisely satisfied only in certain exceptional cases. The preceding formula directly implies that here the probability p_i of all letters of the alphabet must be unity divided by an integral power of the number 2. But in the general case the quantity $-\log p_i$, where p_i is the probability of the i th letter of the alphabet, is, as a rule, not an integer. Hence, the code-word length l_i of the i th letter cannot be equal to $-\log p_i$. However, since in the Shannon-Fano coding method we successively divide our alphabet into groups of *closest*

possible total probability, the code-word length of the i th letter in such coding shall be close to $-\log p_i$. We denote by l_i in this connection the smallest integer not less than $-\log p_i$, i.e., such that

$$-\log p_i \leq l_i < -\log p_i + 1. \quad (\text{A})$$

Inequality (A) can be rewritten as

$$-l_i \leq \log p_i < -(l_i - 1),$$

or

$$\frac{1}{2^{l_i}} \leq p_i < \frac{1}{2^{l_i-1}}. \quad (\text{B})$$

Let us now show that *there exists a coding method in which the code-word length of the i th letter exactly equals this number l_i* . It is just this fact (and not the description of the corresponding coding method)[†] that is needed by us in the proof of the fundamental theorem.

We first show that *in the case of any n numbers l_1, l_2, \dots, l_n , satisfying the inequality*

$$\frac{1}{2^{l_1}} + \frac{1}{2^{l_2}} + \dots + \frac{1}{2^{l_n}} \leq 1, \quad (1)$$

there exists a binary code for which these numbers are the lengths of code words corresponding to n letters of some alphabet. In fact, let n_1, n_2, \dots, n_k be those of the numbers l_1, l_2, \dots, l_n which are, respectively, equal to 1, 2, \dots , k (where $n_1 + n_2 + \dots + n_k = n$, so that k is the maximum value of the numbers l_1, l_2, \dots, l_n). In this case, inequality (1) can be written in the form

$$\frac{n_1}{2} + \frac{n_2}{4} + \frac{n_3}{8} + \dots + \frac{n_k}{2^k} \leq 1.$$

Hence it immediately follows that

$$\begin{aligned} \frac{n_1}{2} &\leq 1, & \text{or } n_1 &\leq 2; \\ \frac{n_2}{4} &\leq 1 - \frac{n_1}{2}, & \text{or } n_2 &\leq 2(2 - n_1); \\ \frac{n_3}{8} &\leq 1 - \frac{n_1}{2} - \frac{n_2}{4}, & \text{or } n_3 &\leq 2[4 - (2n_1 + n_2)]; \\ &\dots\dots\dots \\ \frac{n_k}{2^k} &\leq 1 - \frac{n_1}{2} - \frac{n_2}{4} - \frac{n_3}{8} - \dots - \frac{n_{k-1}}{2^{k-1}}, & \text{or} \\ n_k &\leq 2[2^{k-1} - (2^{k-2}n_1 + 2^{k-3}n_2 + \dots + n_{k-1})] \end{aligned}$$

[†]Regarding this description, see the text in small print on p. 173 et seq.

(see p. 135). It is, however, clear that the condition $n_1 \leq 2$ guarantees the possibility of the choice of n_1 distinct code words of length 1. In analogy to this, the inequality $n_2 \leq 2(2 - n_1)$ indicates the possibility of choosing additionally n_2 code words of length 2 starting with a binary digit other than that which is already 'taken up' by the code words of length 1; as a matter of fact, the number of such 'free' first digits equals $2 - n_1$ and to each of them we can add at the end either the digit 0 or 1. Exactly in the same way, the inequality $n_3 \leq 2[4 - (2n_1 + n_2)]$ allows us to choose additionally n_3 code words of length 3, whose first digit is other than the n_1 digit 'taken up' by the code words of length 1 and the first two digits differ from the n_2 two-digit numbers 'taken up' by the code words of length 2. (In fact, $2n_1 + n_2$ is a number of two-digit binary numbers which either starts with one of the n_1 digits, it being the code word of length 1, or coincides with one of the n_2 code words of length 2, and 4 is the number of all possible two-digit binary numbers with which, in principle, we can start the code word of length 3.) Continuing this reasoning, it is easily seen that the inequality

$$n_k \leq 2[2^{k-1} - (2^{k-2}n_1 + 2^{k-3}n_2 + \dots + n_{k-1})]$$

allows us to choose n_k code words of length k , whose first digit, first two digits, first three digits, . . . , coincide with none of the n_1, n_2, n_3, \dots , code words of length 1, 2, 3, . . . , respectively. In fact, 2^{k-1} is the number of all possible initial combinations of $k - 1$ binary digits and $2^{k-2}n_1 + 2^{k-3}n_2 + \dots + n_{k-1}$ is the number of such combinations that are already 'taken up' (see p. 135). This leads precisely to the conclusion that the fulfilment of inequality (1) assures the possibility of choosing n code words of length l_1, l_2, \dots, l_n satisfying the condition enumerated atop p. 141 in italicized print; these are precisely the code words we can associate with the existing letters of n -letter alphabets.

For completing the existence proof for the required codes, it remains to note only that, by inequality (B) defining the code-word length l_i , we have $1/2^{l_i} \leq p_i$ for all $i = 1, 2, \dots, n$, where p_i is the probability of the i th letter. Thus,

$$\frac{1}{2^{l_1}} + \frac{1}{2^{l_2}} + \dots + \frac{1}{2^{l_n}} \leq p_1 + p_2 + \dots + p_n = 1.$$

Hence the numbers l_1, l_2, \dots, l_i indeed satisfy inequality (1), which is prerequisite for them to be the code-word lengths of a binary code.

The proof of the fundamental coding theorem can now be completed quite easily. In fact, the average number l of binary signals per letter of the original message (in other words, the average code-word length) is, by definition, given by the sum

$$l = p_1 l_1 + p_2 l_2 + \dots + p_n l_n.$$

We now multiply by p_i inequality (A), defining the quantity l_i , sum up all the inequalities so obtained corresponding to the values $i = 1, 2, \dots, n$, and note that

$$H = -p_1 \log p_1 - p_2 \log p_2 - \dots - p_n \log p_n,$$

where $H = H(\alpha)$ is the entropy of the experiment α consisting of determining one letter of the message, and that $p_1 + p_2 + \dots + p_n = 1$. Consequently,

$$H \leq l < H + 1.$$

We now apply this inequality to the case in which the method set forth above is used for coding all possible N -letter blocks (which can be considered as 'letters' of a new alphabet). By virtue of the assumption that successive letters of the message are independent, the entropy of experiments $\alpha_1 \alpha_2 \dots \alpha_N$ considered in the determination of all letters of a block is given by

$$H(\alpha_1 \alpha_2 \dots \alpha_N) = H(\alpha_1) + H(\alpha_2) + \dots + H(\alpha_N) = NH(\alpha) = NH.$$

Consequently, the average code-word length l_N of N -letter blocks satisfies the inequality

$$NH \leq l_N < NH + 1.$$

But in coding N -letter blocks the average number l of binary elementary signals per letter of message is equal to the average code-word length l_N of one block divided by the number N of letters in the block:

$$l = \frac{l_N}{N}.$$

Hence in such coding

$$H \leq l < H + \frac{1}{N},$$

i.e., the average number of elementary signals per letter differs here from the minimum value of H by not more than $1/N$. Letting $N \rightarrow \infty$, we immediately arrive at the fundamental coding theorem.

Before we proceed further, we note that the proof deduced here can be applied also to the more general case in which the successive letters of a text are *mutually dependent*. For this we must rewrite the inequality for the quantity l_N in the form

$$H^{(N)} \leq l_N < H^{(N)} + 1,$$

where

$$\begin{aligned} H^{(N)} &= H(\alpha_1 \alpha_2 \alpha_3 \dots \alpha_N) \\ &= H(\alpha_1) + H_{\alpha_1}(\alpha_2) + H_{\alpha_1 \alpha_2}(\alpha_3) + \dots + H_{\alpha_1 \alpha_2 \dots \alpha_{N-1}}(\alpha_N) \end{aligned}$$

is the entropy of N -letter block which, in the case of the letters of a message being dependent upon each other, is always less than NH (because $H(\alpha_1) = H$ and $H(\alpha_1) > H_{\alpha_1}(\alpha_2) \geq H_{\alpha_1 \alpha_2}(\alpha_3) \geq \dots \geq H_{\alpha_1 \alpha_2 \dots \alpha_{N-1}}(\alpha_N)$). This implies that

$$\frac{H^{(N)}}{N} \leq l \leq \frac{H^{(N)}}{N} + \frac{1}{N},$$

where l is the average number of elementary signals per letter of message. Hence, in this general dependent case, as $N \rightarrow \infty$ (as the block length increases indefinitely) the average number of elementary signals required for the transmission of one letter tends unboundedly to the quantity H_∞ , where

$$H_\infty = \lim_{N \rightarrow \infty} \frac{H^{(N)}}{N}$$

is the 'specific entropy' per letter of a multiletter text (we shall discuss the quantity H_∞ more elaborately later in the next section).†

We now give the *second* proof of our fundamental coding theorem; the successive letters of message are again considered here to be mutually independent. This proof is lengthier than its predecessor, but then it is more instructive since it makes transparent the meaning of the concept of entropy itself (see pp. 55-56). In addition, this new proof shows us that, even in the case of sharply differing probabilities of different letters, when coding very long blocks we can always make use of 'almost uniform' codes by associating with all blocks code words of the same length, except for a certain part of them having a negligibly small probability sum. As regards the latter 'low-probability' blocks, it is easy to understand that they can be coded on an 'as and when occurring' basis: since the probability of the occurrence of any such block is quite small, the method of coding these blocks is of no significant importance.

For greater clarity we start our proof with a detailed examination of the simplest case in which the entire 'alphabet' consists in all of two letters a and b with probabilities $p_1 = p$ and $p_2 = 1 - p = q$. We shall code all possible sequences ('blocks') consisting of N successive letters a and b . The total number of such distinct N -term sequences is 2^N (see pp. 55-56). However, a majority of these

†The existence of the limit H_∞ directly follows from the inequality $H(\alpha_1) \geq H_{\alpha_1}(\alpha_2) \geq H_{\alpha_1 \alpha_2}(\alpha_3) \geq \dots$, which shows that $H(\alpha) = H^{(1)}, (H^{(2)}/2), (H^{(3)}/3), \dots, (H^{(N)}/N), \dots$ is a *monotonically nonincreasing* sequence of positive (i.e., greater than zero) numbers,

N -term sequences have negligible probability. Since the relative frequency of the occurrence of letters a and b is p and q respectively, for a sufficiently large N an aggregate of only those sequences will have a significant probability, in which of the total N numbers of letters the letter a occurs roughly Np times and the letter b occurs the remaining roughly $N - Np = Nq$ times. To be more exact, it can be stated that when N is quite large all sequences to which the relative frequency of occurrence of a is not confined to the range from $p - \epsilon$ to $p + \epsilon$, where ϵ is an arbitrarily chosen very small number (say 0.001, or 0.0001, or 0.000001; for, ϵ can take any of these or even any still smaller number, if only N is sufficiently large), have an extremely small probability sum so that in general they can be ignored in calculation. As to the sequences in which a occurs in the range $N(p - \epsilon)$ to $N(p + \epsilon)$ times, obviously each such sequence also has a small individual probability (for large N the total number of possible sequences is very large, but the probability of each of them individually is quite small), yet the probability sum of all these sequences is quite close to 1.

Let us now note that the number of N -letter sequences, in which a is encountered exactly Np times†, is equal to the number $\binom{N}{Np}$ of combinations of N elements taken Np at a time (i.e., the number of Np -element subsets of a given set of N elements). This makes it necessary to estimate the quantity $\binom{N}{K}$ (see footnote † below) with its dependence on N and K .

In order to make clearer the idea underlying our reasonings, we announce first the derivation (not needed later by us) of the *formula for the number* $\binom{N}{K}$. Suppose that we have N (paper) contours and N different colours, with which we desire to colour these contours—each in its own colour. Since we can paint first contour in any of the N available colours, the second in any of the remaining $N - 1$ colours, the third in any of the $N - 2$ colours not already used, finally the last contour in the only colour left at our disposal, the total number of possible contour colourings is

$$N(N - 1)(N - 2)(N - 3) \dots 1 = N!.$$

Now let us call any K colours to be the ‘first’ and the remaining $N - K$ colours the ‘second’; furthermore, we choose any K contours, which we consider as the ‘first’ (and the other $N - K$ contours as the ‘second’). In such case we have $K!$ ways of painting K ‘first’ contours in the K ‘first’ colours and $(N - K)!$ ways for colouring the remaining $N - K$ contours in the $N - K$ ‘second’ colours.

†If Np is not an integer, then we replace this number by the integer K that is closest to Kp : when N is large the difference between Np and K is negligibly small. A similar observation can also be made in relation to the number $N\epsilon$.

By combining any of the $K!$ ways of painting the K 'first' contours with any of the $(N - K)!$ ways of colouring the remaining contours, we get altogether

$$K! \times (N - K)!$$

ways of colouring N contours in which the chosen K 'first' contours are coloured in K 'first' colours. In addition, since the K 'first' contours can be chosen from the total number N of contours in $\binom{N}{K}$ ways, *the total number of distinct colourings must be*

$$\binom{N}{K} K! (N - K)!.$$

Consequently,

$$N! = \binom{N}{K} K! (N - K)!.$$

implying also the desired equation

$$\binom{N}{K} = \frac{N!}{K!(N - K)!}. \quad (*)$$

The well-known equation (*) gives an *exact* expression for the number $\binom{N}{K}$ in terms of the numbers N and K ; however, for *large* N (and only the case of large N will be of interest to us in the following) it becomes inconvenient. The fact is that $N!$ is the product of N *distinct* factors; an evaluation of its value for large N is rather complicated. Hence, in what follows we shall use not this equation, but an *approximate estimate* of the value of $\binom{N}{K}$. This estimate differs from the right-hand side of (*) mainly in this that it includes only the *powers* of N , K and $N - K$, which are easy to evaluate by taking logarithms. The desired estimate of $\binom{N}{K}$ will be derived below.

Let us consider the same problem of colouring N contours in N colours, which we used for deriving the formula (*), but we do not require now that *each contour be necessarily coloured in its own colour*. In this case, the first contour as before can be coloured in any N colours; however, the second, third, . . . , and last contour can also be coloured in any N colours. Hence, the *total number of colourings* in this case is now given by the expression

$$\underbrace{N \times N \times \dots \times N}_{N \text{ factors}} = N^N.$$

If we now again choose any K 'first' colours and K 'first' contours, then these K contours can be painted in K colours in K^K ways. However, the remaining $N - K$ 'second' contours can be painted in $(N - K)^{N-K}$ ways with $N - K$ 'second' colours. By combining each of the possible K^K paintings of 'first' contours with each of the $(N - K)^{N-K}$ colourings of the remaining contours, we get altogether

$$K^K \times (N - K)^{N-K}$$

different ways of painting all N contours. This number ought to be further multiplied by $\binom{N}{K}$, since $\binom{N}{K}$ is the number of ways in which K 'first' contours can be chosen from the total number of N contours. This yields the number

$$\binom{N}{K} K^K (N - K)^{N-K}$$

of different colourings. However, this number is found to be *not equal* to but *less* than the total number N^N of possible colourings of N contours. In fact, $\binom{N}{K} K^K (N - K)^{N-K}$ is the number of those colourings, in which K 'first' colours are used *exactly* K times (but, there exist also the colourings in which these K colours are used N times say, or are not used at all!). Thus, finally, we get

$$\binom{N}{K} K^K (N - K)^{N-K} < N^N.$$

This also yields the desired estimate of $\binom{N}{K}$ by

$$\binom{N}{K} < \frac{N^N}{K^K (N - K)^{N-K}} \quad (**)$$

Let us now replace K by Np in (**); this converts $N - K$ into $N - Np = N(1 - p) = Nq$. Hence, we get the estimate

$$\binom{N}{Np} < \frac{N^N}{(Np)^{Np} (Nq)^{Nq}} = \frac{N^N}{N^{Np+Nq} p^{Np} q^{Nq}} = \frac{N^N}{N^N p^{Np} q^{Nq}} = \frac{1}{p^{Np} q^{Nq}}$$

for the number $\binom{N}{Np}$ of 'most probable' N -letter sequences of the letters a and b , i.e., the sequences in which the letter a is encountered exactly Np times (and the letter b the remaining $Nq = N - Np$ times). Roughly, there are as many sequences in which a occurs, say, $Np + 1$, $Np + 2$, \dots , $Np + N\epsilon$ times, or

$Np - 1, Np - 2, \dots, Np - N\epsilon$ times as those where a occurs exactly Np times (since in all these cases the deviation of the frequency of occurrence of a from p is very small). Hence, without any risk of serious error, we can consider that the total number of 'probable' sequences (i.e., the sequences such that all other sequences taken together have very small probability, which can be neglected) does not exceed the value

$$M_1 = 2N\epsilon \times \frac{1}{p^{Np}q^{Nq}} = \frac{2N\epsilon}{p^{Np}q^{Nq}},$$

where ϵ is some small number.

We now use the best uniform code for coding M_1 (or less than M_1) probable sequences.† Since the number of such sequences is quite large, the code-word length practically coincides here with the binary logarithm of the number sequences (see, p. 106). Hence this code-word length is not greater than

$$\log M_1 = \log 2\epsilon + \log N - N(p \log p + q \log q).$$

Consequently, the average number of binary digits per letter of message does not exceed here the value

$$\frac{\log M_1}{N} = H + \frac{\log N}{N} + \frac{\log 2\epsilon}{N},$$

where

$$H = -p \log p - q \log q.$$

As $N \rightarrow \infty$ the second and third terms on the right-hand side of the penultimate equation tend to zero (recall that the ratio $\log N/N = -(1/N) \log (1/N)$ tends to zero as $N \rightarrow \infty$; see p. 47). This implies that, if we restrict ourselves to 'probable' sequences, then the average number of binary digits per letter of a message can be made arbitrarily close to H .††

As regards the remaining 'low-probability' sequences, even if we use the number of binary symbols, which is several times greater than H , in coding each letter of these sequences, the average value of the number of such symbols needed per letter of a message remains here all the same almost invariant (since the probability sum of all such sequences is negligibly small). Hence in coding of the remaining sequences it is factually necessary just to take care that none of the corresponding code words coincides with the extension of any other code

†It is easy to see that the application of any nonuniform code to these 'probable sequences' may not confer any substantial advantage. This is due to the fact that probabilities of all such sequences differ only slightly from each other (since the relative frequencies of both the letters are here practically the same in all cases).

††Of course, this number cannot be less than H (see, p. 149).

word being used. This objective can be achieved, for instance, if right from the beginning, we add 1 to the total number of 'probable' sequences (the replacement of M_1 by $M_1 + 1$ obviously does not change any of the above estimates). Then we can make use of the fact that in such a case we certainly have at least one 'free' code word of the same length as all the code words of 'probable' sequences. If we now prefix this 'free' code word to all code words of 'low-probability' sequences, then it will be guaranteed that none of the new words is an extension of one of the old words. After this word, we can add (say) the result of applications to 'low-probability' sequences of any most efficient uniform code, after which finally for all 'low-probability' sequences code words of one and the same length will be obtained, satisfying the required condition.

The general case of an n -letter alphabet, in which individual letters have probabilities p_1, p_2, \dots, p_n respectively, where $p_1 + p_2 + \dots + p_n = 1$, is analyzed almost in the same way. In the case of a long sequence of N letters, the greatest probability will have a sequence in which the first, second, \dots , n th letter is encountered nearly Np_1, Np_2, \dots, Np_n times. The number of sequences in which the first, second, \dots , n th letter occurs exactly Np_1, Np_2, \dots, Np_n times is equal to the number of partitions of a set of N elements into n subsets containing respectively the Np_1, Np_2, \dots, Np_n elements.

Let us now consider the *problem of colouring N contours with N colours such that each colour is used only once*. If we partition the colours into n groups containing, respectively, Np_1, Np_2, \dots, Np_n colours, we can show, in complete analogy with the derivation of equation (*), that the number of such partitions of a set of N elements into n subsets is

$$\frac{N!}{(Np_1)! (Np_2)! \dots (Np_n)!}.$$

This equation generalizes the ordinary equation for the number of combinations $\binom{N}{K}$.† If we consider further the problem of colouring N contours with N colours (as before, partitioned into n groups, of which the first, second, \dots , last contain Np_1, Np_2, \dots, Np_n colours) in which it is *not specified that each colour is used only once*, we can verify in a way similar to the derivation of inequality (**) that the number of partitions of a set of N elements into n subsets we are interested in is less than the number

$$\frac{1}{p_1^{Np_1} p_2^{Np_2} \dots p_n^{Np_n}}.$$

Applying this result to the 'probable' sequences, in which the frequency of

†The derivation of this equation can also be found in [38].

occurrence of the first, second, . . . , n th letters lies, respectively, between $p_1 - \varepsilon$ and $p_1 + \varepsilon$, $p_2 - \varepsilon$ and $p_2 + \varepsilon$, . . . , $p_n - \varepsilon$ and $p_n + \varepsilon$, we find that the total number of such sequences certainly does not exceed the number

$$(2N\varepsilon)^n \times \frac{1}{p_1^{Np_1} p_2^{Np_2} \dots p_n^{Np_n}} = \frac{2^n \varepsilon^n N^n}{p_1^{Np_1} p_2^{Np_2} \dots p_n^{Np_n}}.$$

As to the remaining sequences in which the frequency of occurrence of even one of the letters is not contained within the stated limits, the probability sum of all these sequences is negligibly small, which permits us to ignore them altogether.

It is now just routine to show that by encoding our all 'probable' sequences by means of a most different uniform code we arrive at code words, whose length is less than

$$NH + n \log N + n \log 2\varepsilon,$$

where

$$H = -p_1 \log p_1 - p_2 \log p_2 - \dots - p_n \log p_n.$$

Consequently, the average number of binary symbols required to write one letter does not exceed

$$H + n \frac{\log N}{N} + \frac{n \log 2\varepsilon}{N}.$$

As $N \rightarrow \infty$ this number obviously tends to H . Hence H is equal to the limit average number of binary symbols required per letter of a message in such coding method. This also is just the result we sought to prove.

Finally, it is worth while to reemphasize the main basis of the proof deduced. If we consider all sequences of N letters from an n -letter 'alphabet' (or equivalently, all sequences of N successive outcomes of a many times repeated experiment, which can have n different outcomes), then the total number of such distinct sequences is

$$n^N = 2^{N \log n}.$$

However, the probability of each such individual sequence and even of some appreciable collections of such sequences for large N is completely insignificant. It has been shown that, if we permit ourselves to exclude from consideration a part of the least probable sequences, but only such that the probability sum of all discarded sequences is sufficiently small (say, not exceeding a certain pre-assigned extremely small number δ), then for *any* (arbitrarily small!) δ in the case of N being sufficiently large it is possible to obtain the result that the number

of remaining sequences has the order

$$\left(\frac{1}{p_1}\right)^{Np_1} \left(\frac{1}{p_2}\right)^{Np_2} \dots \left(\frac{1}{p_n}\right)^{Np_n} = 2^{NH},$$

where H is an entropy.[†] Note also the fact that since H is less than $\log n$ (excepting the case in which all letters or all outcomes are equally probable), the number of our 'probable' sequences for extremely large N is incomparably smaller than the total number of all sequences (the ratio

$$2^{NH} : 2^{N \log n} = 2^{-N(\log n - H)}$$

of the number of 'probable' sequences to the number of all sequences rapidly tends to zero as $N \rightarrow \infty$). It has also been shown that for large N it is possible to establish the fact that the relative frequencies of the occurrence of individual letters in our 'probable' sequences differ as little as desired from the most probable frequencies p_1, p_2, \dots, p_n . Since the probability of a sequence depends only on the numbers of individual letters occurring in it (the probability of a sequence in which the first, second, \dots , n th letters occur N_1, N_2, \dots, N_n times is $p_1^{N_1} p_2^{N_2} \dots p_n^{N_n}$), hence it is clear that for large N one can see that all 'probable' sequences differ very little in their probabilities. In other words, we have proved here the statement set in italics on p. 56; this statement determines the main part of the notion of entropy in coding theory.

In view of the specific importance of the the statement brought out, it makes sense to dwell upon it slightly longer and derive one more simple proof of it. In the foregoing, we based our arguments on the calculation of the total number of N -letter sequences in which the frequencies of individual alphabet letters differ little from the corresponding probabilities p_1, p_2, \dots, p_n . In this connection, it was also noted that the probabilities of all such sequences are close to each other and for all practical purposes do not deviate from the probability $p_1^{Np_1} p_2^{Np_2} \dots p_n^{Np_n}$ of sequences in which $N_1 = Np_1, N_2 = Np_2, \dots, N_n = Np_n$,

[†]The phrase 'has the order' implies here that in fact before 2^{NH} there may occur a certain factor proportional to the finite degree of N (that is, proportional to $2^A \log N$, where A is a fixed number); clearly, when N is quite large, this factor is very much less than the basic factor 2^{NH} and does not play an essential role. Note in this connection that in the derivation above we have shown just that the number of 'probable' sequences does not exceed $(2\epsilon)^n n^n 2^{NH}$. It is, however, clear that this number is not less than the number of sequences containing the first, second, \dots , n th letter exactly Np_1, Np_2, \dots, Np_n times. It has been shown above that the last number is necessarily greater than

$$\frac{1}{p_1^{Np_1} p_2^{Np_2} \dots p_n^{Np_n}} = 2^{NH}.$$

Thus, to within a factor of the order of the finite degree of N the number of 'probable' sequences coincides with the number 2^{NH} .

i.e., the frequencies of occurrence of each of the n alphabet letters precisely coincide with the probabilities p_1, p_2, \dots, p_n . The preceding probability can obviously be rewritten in the form

$$(2^{\log p_1})^{Np_1} (2^{\log p_2})^{Np_2} \dots (2^{\log p_n})^{Np_n} = 2^{N(p_1 \log p_1 + p_2 \log p_2 + \dots + p_n \log p_n)} = 2^{-HN}.$$

Since $H = -p_1 \log p_1 - p_2 \log p_2 - \dots - p_n \log p_n$ is a fixed finite number and N is very large, it is clear that the probability 2^{-HN} is quite small. Let us now note that the formula derived immediately implies the estimate required by us of the total number of different 'probable' sequences. In fact, the probability sum of all such sequences is quite close to unity (it differs from unity by just some extremely small number); since the probability of the sum of incompatible events is equal to the sum of the corresponding probabilities, it is clear that the total number of the considered sequences must be close to unity divided by the probabilities of individual sequences, i.e., close to the number 2^{HN} . Thus, the statement we are interested in is proved if we can just show that in a collection of all n^N possible N -letter sequences it is possible to discard some collection of 'low-probability' sequences (whose probability sum for sufficiently large N can be made as small as desired) so that all the remaining sequences have practically the same probability 2^{-NH} .

Now we can easily evaluate the probability of any sequence of N letters of an n -letter alphabet (where the probabilities of first, second, \dots , n th letters are, respectively, p_1, p_2, \dots, p_n), if these sequences are such that N letters are chosen successively one after the other independently of those chosen previously. This probability obviously equals the product $p_{i_1} p_{i_2} \dots p_{i_N}$, where i_1, i_2, \dots, i_N are the numbers of successive letters of our sequence. Consequently, the logarithm of this probability is given by the relation

$$\log p_{i_1} + \log p_{i_2} + \dots + \log p_{i_N} = \frac{\log p_{i_1} + \log p_{i_2} + \dots + \log p_{i_N}}{N} \times N.$$

But the variables $\log p_{i_1}, \log p_{i_2}, \dots, \log p_{i_N}$ are all defined by the results of experiments consisting of the choice of one of the n alphabet letters. Hence these are all *random variables*, which can take n values $\log p_1, \log p_2, \dots, \log p_n$ with probabilities p_1, p_2, \dots, p_n , respectively. By applying the law of large numbers proved on pp. 34-36 to such a random variable, we find that with a probability, which for sufficiently large N can be considered as arbitrarily close to unity, the arithmetic mean

$$\frac{\log p_{i_1} + \log p_{i_2} + \dots + \log p_{i_N}}{N}$$

differs from

$$\text{m.v. } \log p = p_1 \log p_1 + p_2 \log p_2 + \dots + p_n \log p_n = -H$$

by not more than a given very small number ϵ . But this also implies that *among the number of all N -letter sequences it is possible to disregard some collection of 'low-probability' sequences of very small probability sum such that the probability of all the rest of the sequences remains roughly the same and extremely close to 2^{-HN}* . The last statement is also the one which we desired to prove.

Let us further sketch briefly the role of the assumption specifying that the successive letters of a message are chosen each time *independently* of all the preceding letters. On pp. 161-62 it has been shown that first proof of the fundamental coding theorem considered does not in fact depend on the fulfilment of this condition. However, in the general case of *mutually dependent* letters the value of the entropy H of one letter must be replaced by the per letter *specific entropy*

$$H_\infty = \lim_{N \rightarrow \infty} \frac{H^{(N)}}{N} \quad (\text{where } H^{(N)} \text{ is the entropy of a block of } N \text{ letters}).$$

Starting from this it seems natural to expect that the second proof must in fact be applicable also to the general case of a message with mutually dependent letters, although in the course of this proof the assumption of independence of the letters of a message is essentially used. In other words, it seems natural to expect that even in the case of a message whose letters depend on each other *among all N -letter sequences, where N is sufficiently large, one can extract a collection of 'probable' sequences, whose probability sum differs very little from unity, the number of these probable sequences being of the order $2^{H_\infty N} \approx 2^{H^{(N)}}$ and the probability of each of them being close to $2^{-H_\infty N} \approx 2^{-H^{(N)}}$* . The statement set in italics occupies a very important place in information theory; however, its proof is not quite straightforward and, moreover, it cannot be obtained in general for all cases without exception since it demands that the probability distributions for successive letters of a message satisfy certain additional conditions. (These additional conditions are of a great variety and are always accomplished in practice, but even their formulation entails the introduction of several quite new and nonelementary probabilistic notions.) Note also that these additional conditions can be chosen in different ways: thus, for one such condition the statement made above was proved by Shannon ([21]), Theorem 3), while later on entirely different, quite general conditions for its validity were specified by McMillan [68]. We shall not further elaborate on this aspect and, instead, we refer the reader to [8], [9], [11] and [23], in which the subject is analyzed in great details.

All the preceding arguments of this section easily carry over also to the case of m -ary codes employing m elementary signals. Thus, (say) for constructing m -ary Shannon-Fano codes it is required only to partition groups of symbols not into two but into m parts of closest possible probability. Similarly, for constructing m -ary Huffman code it is necessary to use the contraction operation of the alphabet, in which each time we combine not two but m letters of the original alphabet, having the lowest probabilities. In view of the importance of the Huffman code, we deal with the last question in slightly more detail. The

contraction of an alphabet, in which m letters are replaced by one, clearly reduces the number of letters by $m - 1$. The obvious prescription for the construction of m -ary codes is that the sequence of 'contractions' finally leads us to an alphabet of m letters (associated with m code signals), and hence it is necessary that the number n of original alphabet letters be represented in the form

$$n = m + k(m - 1),$$

where k is an integer. This, however, can always be achieved by adding, if required, to the original alphabet a few 'fictitious letters', whose probabilities are considered to be zero. Then the construction of an m -ary Huffman code and the proof of its optimality (among all m -ary codes) are carried out in exactly the same way as in the case of a binary code. Thus, for instance, in the case of the 6-letter alphabet considered above, having the probabilities 0.4, 0.2, 0.2, 0.1, 0.05 and 0.05, for the construction of a *ternary* Huffman code it is required to affix to our alphabet one additional fictitious letter of zero probability and act further as indicated in the accompanying table.

No. of letters	Probabilities and code words					
	Original alphabet			Contracted alphabets		
1	0.4	0		0.4	0	
2	0.2	2		0.2	2	
3	0.2	10		0.2	10	
4	0.1	11		0.1	11	
5	0.05	120		0.1	12	
6	0.05	121				
7	0					

Both proofs of the fundamental coding theorem deduced above carry over to the case of m -ary codes in a straightforward manner. In particular, the corresponding modification of the first proof is based on the fact that *any n numbers l_1, l_2, \dots, l_n , satisfying the inequality*

$$\frac{1}{m^{l_1}} + \frac{1}{m^{l_2}} + \dots + \frac{1}{m^{l_n}} \leq 1, \quad (2)$$

form the code-word lengths of some m -ary code for an n -letter alphabet. The proof of this fact is precisely a reiteration of the arguments deduced on pp. 159-60 for the case of $m = 2$; hence, we need not dwell upon it here. Using inequality (2) in the same way as inequality (1) on p. 159, it is easy to obtain the following result (called the *fundamental coding theorem for m -ary codes*): *in any coding method, using an m -ary code, the average number of elementary signals per letter of a message can never be less than the ratio $H/\log m$ (where H is*

the entropy of one letter of the message); however, the former can always be made as close as desired to the latter quantity, if sufficiently long N -letter 'blocks' are coded directly and not the letters. Hence, it is clear that if L elementary signals (taking m distinct values) can be transmitted through a communication channel in unit time, then the information transmission rate over such a channel cannot be greater than

$$v = \frac{L \log m}{H} \text{ letters/unit time;}$$

the transmission at a rate as close as desired to v (but less than v !) is possible, however. The variable

$$C = L \log m,$$

appearing in the numerator of the expression for v , depends only on the communication channel itself (while the denominator H characterizes the message to be transmitted). This variable defines the greatest amount of information units that can be transmitted over our channel in unit time (because one elementary signal, as we know, can contain at most $\log m$ units of information); it is called the *channel capacity*. The notion of channel capacity occupies an important place in communication theory; we shall come back to this later also (see Sections 4.3.6 (pp. 246-51) and 4.4).

We offer one more remark related to the first proof of the fundamental coding theorem derived on p. 158 et seq. The fact of the existence of a binary code plays a central role in this proof, in which the code-word length l_i of the i th letter satisfies the inequalities

$$-\log p_i \leq l_i < -\log p_i + 1, \quad (\text{A})$$

or, equivalently,

$$\frac{1}{2^{l_i}} \leq p_i < \frac{1}{2^{l_i-1}}. \quad (\text{B})$$

In the case of an arbitrary m -ary code these inequalities assume the form

$$-\frac{\log p_i}{\log m} \leq l_i < -\frac{\log p_i}{\log m} + 1, \quad (\text{A}')$$

or, equivalently,

$$\frac{1}{m^{l_i}} \leq p_i < \frac{1}{m^{l_i-1}}. \quad (\text{B}')$$

The existence of a binary code satisfying (A) and (B) is proved above, relying on inequality (1) on p. 159 but the explicit expressions of the code words are not set out in the proof. In the case of an m -ary code, in exactly the same way, inequality (2) on p. 172 can be used. We now describe a method for the explicit construction of the corresponding code words. For

simplicity, we shall confine ourselves here to the case of a *decimal* code, associating some sequences of *digits* 0, 1, . . . , 9 with each of the n -alphabet *letters*.† For such a decimal code, the inequalities (A') and (B') obviously assume the form

$$-\lg p_i \leq l_i < -\lg p_i + 1 \quad (\text{A''})$$

(where \lg indicates common decimal logarithms!), and

$$\frac{1}{10^{l_i}} \leq p_i < \frac{1}{10^{l_i-1}}. \quad (\text{B''})$$

Arrange the whole 'alphabet' in the order of decreasing probabilities $p_1 \geq p_2 \geq p_3 \geq \dots \geq p_n$. Among these probabilities, we may obviously encounter even identical ones; hence the probability by itself cannot uniquely characterize the corresponding letters. If, however, we set up the sums

$$P_1 = 0, \quad P_2 = p_1, \quad P_3 = p_1 + p_2, \quad P_4 = p_1 + p_2 + p_3, \dots, \quad P_n = p_1 + p_2 + \dots + p_{n-1},$$

then these sums are plainly all distinct. Thus, the n numbers P_1, P_2, \dots, P_n can be considered as a distinctive 'alphabet', corresponding uniquely to the original n -letter alphabet. We are now required only to encode the new 'alphabet', i.e., to associate a definite sequence of elementary signals (or digits) with each of the n numbers P_i . Such coding solves also simultaneously the problem of coding the original alphabet.

It is not difficult to indicate a method for solving the problem of coding the number set P_1, P_2, \dots, P_n . Let us represent each of the numbers P_i (less than unity!) in the form of a (in general an infinite) *decimal fraction*:

$$P_i = 0.a_1a_2a_3 \dots a_k \dots,$$

where a_1, a_2, a_3, \dots are any digits (if P_i is expressed in the form of a *finite* decimal fraction, then all digits a_k from a certain digit onwards are zero). Every P_i is, in turn, associated with the infinite sequence $a_1a_2a_3 \dots$ of digits (i.e., of elementary signals); here the n sequences of digits so obtained are obviously all distinct because no two P_i are equal to each other.

Now note that the distinction between the introduced sequences $a_1a_2a_3 \dots$ cannot be manifested only in digits which are quite far away from the initial digit. In fact, it is obvious that

$$P_{i+1} - P_i = p_i, \quad P_{i+2} - P_i = p_i + p_{i+1}, \dots$$

Hence, by inequality (B'') all numbers $P_{i+1}, P_{i+2}, \dots, P_n$ differ from P_i by not less than $1/10^{l_i}$ and therefore the expansions of all these numbers into decimal fraction differ from the

†The general case differs from this mainly in that it entails the expansion of the numbers P_i appearing below into an (infinite) *m-ary fraction*, i.e., the representation of each number P_i in form of the sum

$$P_i = \frac{a_1}{m} + \frac{a_2}{m^2} + \frac{a_3}{m^3} + \dots + \frac{a_k}{m^k} + \dots,$$

where all 'digits' $a_1, a_2, \dots, a_k, \dots$ in the formation of this fraction assume any of the values 0, 1, . . . , $m - 1$. We recommend that the reader undertake the related construction as an independent exercise.

decimal fraction expansion of the number P_i in the l_i th, or even preceding to l_i th, digit. In other words, *all decimal fractions of $P_{i+1}, P_{i+2}, \dots, P_n$ differ from the decimal fraction of P_i in at least one of the first l_i digits.* Hence, if we leave out just the first l_i digits in the decimal expansion of P_i (where $i = 1, 2, \dots, n$), then we obtain n (finite!) decimal fractions, which are all distinct and none of which is a prefix of the other. The corresponding n sequences $a_1 a_2 a_3 \dots a_{l_i}$ of digits (associated with the n letters of the original alphabet) form the required decimal code.

It is shown above that *any n numbers l_1, l_2, \dots, l_n such that*

$$\frac{1}{m^{l_1}} + \frac{1}{m^{l_2}} + \dots + \frac{1}{m^{l_n}} \leq 1, \quad (2)$$

form the code-word lengths of some m -ary code, which associates n letters of the alphabet with n sequences of elementary signals, taking m possible values. Setting the corresponding arguments in the converse order, it is straightforward to show also that the *code-word lengths l_1, l_2, \dots, l_n of any m -ary code for an n -letter alphabet necessarily satisfy inequality (2).* This has already been factually established at the end of the previous chapter (see pp. 135-36), albeit without using the terminology of this chapter. Thus, *it is necessary and sufficient that inequality (2) be satisfied in order that the numbers l_1, l_2, \dots, l_n be able to form the code-word lengths of some m -ary code.* This statement was first proved in 1949 by the American scientist Kraft in his unpublished dissertation (see, for example [8] and [1]), and later it was further extended by McMillan [69]; hence, inequality (2) is often called the *Kraft inequality* or *McMillan inequality*. The generalization due to McMillan is connected with the circumstance that so far we have considered only codes satisfying the general condition set in italics atop p. 141 (and termed them *instantaneous* or *instantaneously decodable* in the footnote on the same page); it is only to these codes that all the arguments deduced above are related. McMillan has shown, however, that *condition (2) is necessary and sufficient also for the existence of a uniquely decipherable* (but not necessarily instantaneous) *m -ary code with code-word lengths l_1, l_2, \dots, l_n .* Since any instantaneous code is at the same time also uniquely decipherable, it is obviously required to prove only the *necessity* of the stated inequality for any uniquely decipherable code, i.e., the fact that in the case of any uniquely decipherable m -ary code for an n -letter alphabet the *code-word lengths l_1, l_2, \dots, l_n necessarily satisfy inequality (2).* The last statement has been proved in a most straightforward manner by Karush [67], whose proof we shall also follow in our presentation.

Denote by A the sum

$$\frac{1}{m^{l_1}} + \frac{1}{m^{l_2}} + \dots + \frac{1}{m^{l_n}},$$

where l_1, l_2, \dots, l_n are code-word lengths of some uniquely decipherable m -ary code associated with the n -letter alphabet. Let us now set up the expression

$$\begin{aligned} A^t &= \left(\frac{1}{m^{l_1}} + \frac{1}{m^{l_2}} + \dots + \frac{1}{m^{l_n}} \right)^t \\ &= \underbrace{\left(\frac{1}{m^{l_1}} + \frac{1}{m^{l_2}} + \dots + \frac{1}{m^{l_n}} \right)}_{t \text{ times}} \underbrace{\left(\frac{1}{m^{l_1}} + \frac{1}{m^{l_2}} + \dots + \frac{1}{m^{l_n}} \right)}_{t \text{ times}} \\ &\quad \dots \underbrace{\left(\frac{1}{m^{l_1}} + \frac{1}{m^{l_2}} + \dots + \frac{1}{m^{l_n}} \right)}_{t \text{ times}}. \end{aligned}$$

Removing the parantheses in the last product, we obtain the sum of n^t terms of the form $1/m^N$, where each exponent N is equal to one of the sums of the form $l_{i_1} + l_{i_2} + \dots + l_{i_t}$. The numbers i_1, i_2, \dots, i_t take here the values $1, 2, \dots, n$ and of course they need not be all distinct. If it is assumed that the lengths of n code-words for a uniquely decipherable m -ary code are so ordered that $1 \leq l_1 \leq l_2 \leq \dots \leq l_n$, then the two inequalities

$$t \leq N \leq tl_n$$

hold for every sum

$$N = l_{i_1} + l_{i_2} + \dots + l_{i_t}.$$

In fact, it is clear that $N = t$ if $l_{i_1} = l_{i_2} = \dots = l_{i_t} = 1$, and $N = tl_n$ if $l_{i_1} = l_{i_2} = \dots = l_{i_t} = l_n$. Now denote by K_N the number of distinct sums $l_{i_1} + l_{i_2} + \dots + l_{i_t}$, taking the value N . It is then easy to see that by removing the parantheses in the expression for A^t , we get

$$A^t = \left(\frac{1}{m^{l_1}} + \frac{1}{m^{l_2}} + \dots + \frac{1}{m^{l_n}} \right)^t = K_t \frac{1}{m^t} + K_{t+1} \frac{1}{m^{t+1}} + \dots + K_{tl_n} \frac{1}{m^{tl_n}},$$

where, of course, some of the coefficients $K_t, K_{t+1}, \dots, K_{tl_n}$ can take zero values. Now note that the number K_N of distinct sums $l_{i_1} + l_{i_2} + \dots + l_{i_t}$, taking the value N , is equal to the number of distinct t -letter words $b_{i_1}b_{i_2} \dots b_{i_t}$ (where b_1, b_2, \dots, b_n are our alphabet letters) to be encoded by a sequence of N elementary signals. It is easy to show that

$$K_N \leq m^N$$

for any uniquely decipherable code. Indeed, m^N is the total number of distinct sequences of N signals, each of which can take one of the m values, and if any two distinct words were encoded by the same sequence of elementary signals, then this would imply that the code is not uniquely decipherable. Hence for any (natural) t

$$\begin{aligned} A^t &= K_t \frac{1}{m^t} + K_{t+1} \frac{1}{m^{t+1}} + \dots + K_{tl_n} \frac{1}{m^{tl_n}} \\ &\leq m^t \frac{1}{m^t} + m^{t+1} \frac{1}{m^{t+1}} + \dots + m^{tl_n} \frac{1}{m^{tl_n}} = tl_n - (t-1) \leq tl_n. \end{aligned}$$

But this also implies that

$$A \leq 1$$

(i.e., the inequality (2) holds!). In fact, for any $A > 1$ the variable A^t increases with increasing t faster than ct , where c is an arbitrary fixed number† (say, l_n), and hence for sufficiently large t the inequality $A^t > l_n t$ is necessarily satisfied.

From the fact that for both the instantaneous code and any uniquely decipherable code the necessary and sufficient conditions for the existence of a code with given code-word lengths

†Denote $1/t$ by p ; then, $\log(A^t) = t \log A = \log A/p$, and

$$\log(ct) = \log c + \log t = \log c - \log p.$$

It is clear that when p is small (i.e., when t is large) the first of these numbers is considerably greater than the second, because $\log c$ is a constant number (independent of p), $\log A > 0$ (since $A > 1$), but the ratio $(-\log p) : [(\log A)/p] = (1/\log A)(-p \log p)$ vanishes as $p \rightarrow 0$ (see p. 47).

l_1, l_2, \dots, l_n has one and the same form (2), it follows that *for any uniquely decipherable m -ary code there exists an instantaneous code having the same code-word length of letters as in the case of the original uniquely decipherable code.* But this in turn implies in particular that *Huffman codes are optimal* (i.e., have the least average code-word length per letter) not only among all instantaneous codes (this fact is shown on pp. 156-158; see, also p. 172), but also in general *among all uniquely decipherable codes.*

4.3. Entropy and information of various messages encountered in practice

The preceding two sections were devoted to the problem of the coding and transmission of an abstract 'message' written in any 'language' whose alphabet consists of n letters. We shall now discuss the conclusions that can be derived in relation to specific types of messages used in human communications—in the first place messages expressed in the *English* language or in some foreign languages. There exists extensive literature on this subject (see, for example, [1], [5], [6], [17], [147], [148], [173] and [174], which will be reviewed only partly below).

4.3.1. Written Language

The basic result of Sec. 4.1 is related to an M -letter message transmission (where M is sufficiently large) over a communication channel admitting m distinct elementary signals. The result states that for such transmission it is necessary to send not less than $M \log n / \log m$ signals, where n is the number of different 'alphabet' letters by means of which the message is written; moreover, there exists a coding method which enables us to approach as closely as desired the indicated bound of $M \log n / \log m$ signals. Since the *English* 'telegraphic' alphabet contains 27 letters (the 26 ordinary *English* letters and also the 'zero letter'—the space between words), hence for the transmission of an M -letter message composed of *English* words, it is necessary to send

$$M \frac{\log 27}{\log m} = M \frac{H_0}{\log m}$$

elementary signals. Here

$$H_0 = \log 27 \approx 4.75 \text{ bits}$$

is the entropy of an experiment that consists of receiving one letter of the *English* text (the information contained in one letter), subject to the condition that all letters are considered *equally probable*.

In real life, however, the appearance of different letters in an *English* language text is far from being equally probable. Thus, for instance, in any text the letters *E* and *T* occur more frequently than *Q* or *J*; since the average word-length in the *English* language is considerably less than 26 letters, the probability of the

occurrence of a space ('zero letter') by far exceeds the value $1/27$, which we would have obtained if all 27 letters were equally probable. Hence, the information contained in one letter of any intelligible *English* text is always less than $\log 27$ ($\approx 4\frac{1}{2}$ bits). This implies that it is impossible to produce a text composed of *English* letters, in which each letter contains $\log 27$ bits of information by just taking an excerpt from some *English* book. To achieve this, it is necessary to write 27 letters on separate cards; place all these cards in an urn and then draw them one by one, each time writing down the letter drawn and replacing the card in the urn and mixing well the contents of the urn. Carrying out such an experiment, we arrive at a 'sentence' which looks as follows (cf. Shannon [21]):

XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGHYD
QPAAMKBZAACIBZLHJQD

This text may be called a 'zero-order letter approximation' to *English*. Though it is made up of *English* letters, it has obviously little in common with the *English* language.

For a more accurate calculation of the information contained in one letter of an *English* text, it is necessary to know the *probabilities* of the occurrence of different letters. These probabilities can be determined approximately by taking a sufficiently large excerpt written in *English* and calculating for it the relative frequencies of individual letters. Strictly speaking, these frequencies may depend also upon the character of the text and the singularities of the style of the individual author. For example, it seems plausible that in some scientific books the frequencies of individual letters undergo a change due to the appearance of many special terms and foreign words. Even greater deviations from usual letter frequencies can be found sometimes in poetry†, or in some refined fiction work. A striking example of the latter is provided in [17, Chap. 3]: it is related to the 267-page novel *Gadsby* by the American author Ernest Vincent Wright published in 1939, which does not contain anywhere the letter *E* (ordinarily the most frequently used letter in *English* alphabet!). Some other peculiar examples of this sort related to the *German* and *Portuguese* literature are also listed in [17]. Hence for the reliable determination of the 'average frequency' of letters it is desirable to have a collection of different texts taken from different sources. As a rule, however, the deviation from the 'normal letter frequencies' is nevertheless comparatively small and it can be ignored in a first approximation. Approximate values of the frequencies of individual *English* letters are listed in the accompanying table (see, for example, Shannon [159], Quastler in [19], Pierce

†Let us, for instance, mention the poem 'Rush' by the Russian poet K. D. Balmant. In this poem, the rustling of rushes is described by the repeated appearances of the (usually quite infrequent) Russian hissing letters ш (sh) and ч (ch).

[17], Abramson [1], Reza [152], Pratt [149], which contain slightly different numerical data; the space between words is denoted here by a dash).

TABLE

<i>Letter</i>	<i>Relative frequency</i>	<i>Letter</i>	<i>Relative frequency</i>
—	0.182	<i>m</i>	0.021
<i>e</i>	0.107	<i>u</i>	0.020
<i>t</i>	0.086	<i>g</i>	0.016
<i>a</i>	0.067	<i>y</i>	0.016
<i>o</i>	0.065	<i>p</i>	0.016
<i>n</i>	0.058	<i>w</i>	0.013
<i>r</i>	0.056	<i>b</i>	0.012
<i>i</i>	0.052	<i>v</i>	0.007
<i>s</i>	0.050	<i>k</i>	0.003
<i>h</i>	0.043	<i>x</i>	0.001
<i>d</i>	0.031	<i>j</i>	0.001
<i>l</i>	0.028	<i>q</i>	0.001
<i>f</i>	0.024	<i>z</i>	0.001
<i>c</i>	0.023		

Equating these frequencies to the probabilities of the occurrence of the corresponding letters, the approximate value† of the entropy of the *English* text letter is given by

$$\begin{aligned}
 H_1 = H(a_1) &= -0.182 \log 0.182 - 0.107 \log 0.107 - 0.086 \log 0.086 \\
 &\quad - \dots - 0.001 \log 0.001 \\
 &\approx 4.03 \text{ bits.}
 \end{aligned}$$

From a comparison of this value with $H_0 = \log 27 \approx 4.75$ bits it is seen that the irregularity in the occurrence of different letters of the alphabet leads to a

†Since the values of the frequencies of individual letters in an excerpt containing a finite number N of letters do not coincide with the corresponding probabilities, it is clear that the value of the entropy obtained by substituting probabilities for frequencies is not exact. An estimation of the accuracy of the values of the entropies thus obtained and corrections involved to these values when N is not large enough have been considered, for instance, by Basharin in [74] and Miller in [19, pp. 95-100]. See also Blyth [79], Pfaffelhuber [144] and Nemetz [133].

reduction in the information contained in one *English* text letter by roughly 0.72 bit.

Making use of this fact, we can reduce the number of elementary signals required for the transmission of an *English* M -letter message to the value $M(H_1/\log m)$ (i.e., in the case of a binary code to the value $H_1 M \approx 4.03 M$). A reduction in the required number of elementary signals can be achieved by coding individual *English* alphabet letters by the Shannon-Fano method (see p. 150 et seq.). It is not difficult to verify that the application of this method leads to the accompanying table of code words.

TABLE

<i>Letter</i>	<i>Code word</i>	<i>Letter</i>	<i>Code word</i>	<i>Letter</i>	<i>Code word</i>
—	111	<i>i</i>	0101	<i>r</i>	0110
<i>a</i>	1001	<i>j</i>	0000000010	<i>s</i>	01001
<i>b</i>	000001	<i>k</i>	00000001	<i>t</i>	101
<i>c</i>	001001	<i>l</i>	00110	<i>u</i>	00011
<i>d</i>	00111	<i>m</i>	001000	<i>v</i>	0000001
<i>e</i>	110	<i>n</i>	0111	<i>w</i>	000010
<i>f</i>	00101	<i>o</i>	1000	<i>x</i>	0000000011
<i>g</i>	000101	<i>p</i>	000011	<i>y</i>	000100
<i>h</i>	01000	<i>q</i>	0000000001	<i>z</i>	0000000000

The average number of elementary signals required for the transmission of one letter of a message under such a coding method is given by

$$0.375 \times 3 + 0.298 \times 4 + 0.196 \times 5 + 0.117 \times 6 + 0.007 \times 7 + 0.003 \times 8 + 0.004 \times 10 \approx 4.11,$$

i.e., it is considerably lower than $H_0 \approx 4.75$ and does not differ sharply from $H_1 \approx 4.03$.†

†Besides, it is rather difficult to decipher a message encoded by such a method, and this renders this code of little practical value. The difficulty in deciphering can be verified, for instance, by attempting to decode the following 'sentence':

101010001101110100101000100101110111100001111100101100101111000111001001100000
 1111101110101010011110011101010010100101010010010001100110101111011000
 11100111110001001100000111110

(Decoding is facilitated appreciably if we set up beforehand all code words in the order of decreasing probabilities of corresponding letters.)

The average number of elementary signals per letter of a message to be transmitted, even when it equals the value $H_1/\log m$, is not the best, however. In fact, in defining the entropy $H_1 = H(\alpha_1)$ of experiment α_1 , consisting of determining one letter of an *English* text, we had considered all letters to be *independent*. This means that for making up a 'text' in which every letter contains $H_1 \approx 4.03$ bits of information, we must use an urn containing 1,000 well-mixed tickets, of which nothing is written on 182, the letter *e* is written on 107, on 86 the letter *t*, . . . , and, finally, on 1 ticket the letter *z* is written (see the frequency table of *English* letters on p. 179). By drawing the tickets from this urn one by one we may arrive at a 'sentence' that looks like the following :†

OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI
ALHENHTTPA OOBTTVA NAN BRL

This 'first-order letter approximation' to *English* is somewhat more akin to intelligible written *English* than its predecessor (we observe here at least a plausible distribution of the number of vowels and consonants and the mean word length is close to the average word length of *English* language), but it is obviously still far from being a reasonable text.

The dissimilarity of our sentence from an intelligible text is naturally explained by the fact that in real life the successive letters of an *English* text are not at all independent of each other. Thus, for example, the letter *Q* in *English* is always followed by *U* (so that the combinations *QA*, *QB* or *QX*, for example, have a zero probability); *T* is most often followed by *H* (*TH* occurs most frequently of all two-letter combinations or digrams in the *English* language); similarly, the letters *O* and *W* are most often followed by *R* and *E*, respectively; the probability of the occurrence of a vowel after a consonant is significantly higher than the probability of its occurrence after another vowel, and so on. The existence of such auxiliary regularities in the *English* language, for which no allowance is made in our 'sentence', leads to a further reduction in the amount of uncertainty (entropy) of one letter of the *English* text. Hence, in the transmission of such a text over a communication channel, we can still reduce the average number of elementary signals required to transmit one letter. It is not difficult to comprehend how this reduction can be characterized numerically. For this it is necessary only to calculate the *conditional entropy* $H_2 = H_{\alpha_1}(\alpha_2)$ of experiment α_2 that consists of the determination of one letter of the *English* text, given that we know the outcome of experiment α_1 that consists of the determination of the *preceding letter* of the same text. (Note that when the next letter of a message is received, we always know already the preceding letter.) By what has been stated on pp. 62-63, the conditional entropy H_2 is defined by the formula

†See Shannon [21] (cf. also Dobrushin [91]). As explained in these papers, instead of drawing from an urn with 1,000 tickets we can undertake a considerably easier procedure, viz. we take any *English* book and choose from it a series of letters *at random*.

$$\begin{aligned}
H_2 &= H_{\alpha_1}(\alpha_2) = H(\alpha_1\alpha_2) - H(\alpha_1) \\
&= -p(- -) \log p(- -) - p(- a) \log p(- a) \\
&\quad - p(- b) \log p(- b) - \dots - p(zz) \log p(zz) \\
&\quad + p(-) \log p(-) + p(a) \log p(a) \\
&\quad + p(b) \log p(b) + \dots + p(z) \log p(z),
\end{aligned}$$

where we denote by $p(-)$, $p(a)$, $p(b)$, \dots , $p(z)$ the probabilities (frequencies of individual letters of the *English* language (their values are indicated on p. 179), and by $p(- -)$, $p(- a)$, $p(- b)$, \dots , $p(zz)$ the probabilities (frequencies) of all possible digrams, i.e., two-letter combinations. Probability tables of such digrams in *English* texts, computed for the purpose of cryptanalysis (i.e., for deciphering the encoded messages), are available (see for example, Pratt [149]). For an approximate determination of such 'digram probabilities' it is only necessary to calculate the frequencies of the appearance of different combinations of two adjoining letters in any sufficiently long *English* excerpt; in doing so, it is obviously possible to assume in advance that the probabilities $p(- -)$, $p(qa)$ and a series of others (say, $p(xx)$, $p(jj)$, $p(qx)$ and so on) are zero. The numerical value of H_2 will be given later in this section. Here we only emphasize that, by virtue of the results of Section 2.2, we can be convinced of the fact that the value of the conditional entropy $H_2 = H_{\alpha_1}(\alpha_2)$ is *less* than that of the unconditional entropy H_1 .

The quantity H_2 can be described as the 'average information' contained in the definition of the outcome of the following experiment. Let us assume that there are 27 urns to denote 27 letters of the *English* alphabet and that each urn contains tickets on which are written different digrams (i.e., two-letter combinations) starting with the letter denoting the urn. Suppose that the number of tickets in the urn with a specific digram is proportional to the frequency (probability) of the corresponding digram. The experiment consists of repeatedly drawing tickets from the urns and writing out the last letters obtained from them. In this process, each time (starting with the second one) the tickets are drawn from that urn which contains the digram beginning with the last letter written out; after the letter is noted, the ticket is replaced in the urn from which it was drawn and the urn contents are thoroughly shuffled. (Instead of urns, it is also possible to make use of any *English* book, starting each time with a randomly chosen place, to seek the first appearance of the last letter chosen by us and add the letter that follows it to the already existing text. It is clear that such book experiment is much easier to perform than the corresponding urn experiment.) An experiment of this sort leads to a 'sentence' that looks like the following:

ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY
 ACHIN D ILONASIVE TUCOOWE AT TEASONARE FUSO
 TIZIN ANDY TOBE SEACE CTISBE

Phonetically, this 'second-order letter approximation' is appreciably closer to the *English* language than its predecessor set forth on p. 181 (for instance, here we have not only a plausible correlation between the numbers of vowel and consonant letters but are also nearer to their usual alternation sequence, because of which the sentence can be 'pronounced', though not without difficulty). We may also point out that there are several 'genuine' *English* words in this sentence (e.g., *ON*, *ARE*, *BE*, *AT*), but the previous example contains no such word.

Thus, it is obvious that the quantity $H_2/\log m$ also fails to yield the best possible estimate of the minimum value of the average number of elementary signals required for the transmission of one letter of the *English* text. The fact is that in the *English* language (and for that matter in any other language) each letter depends not only on a letter that immediately precedes it but also on a *series* of preceding letters. For instance, it is known that the three-letter combination (trigram) *THE* quite frequently appears in the *English* language (and even a five-letter combination—*THE*—is rather probable), but the trigram *THH* is practically impossible. It is also known that after two consonants a vowel follows much more frequently than a third consonant and after two vowels a consonant is almost obligatory and so on. Hence, the knowledge of *two* preceding letters reduces still further the uncertainty of the event consisting of the determination of the succeeding letter, which is revealed in the difference $H_2 - H_3$ being positive, where H_3 is the 'conditional second-order entropy' defined by

$$\begin{aligned} H_3 &= H_{\alpha_1\alpha_2}(\alpha_3) = H(\alpha_1\alpha_2\alpha_3) - H(\alpha_1\alpha_2) \\ &= -p(---) \log p(---) - p(-a) \log p(-a) - \dots - p(zzz) \log p(zzz) \\ &\quad + p(-) \log p(-) + p(-a) \log p(-a) + \dots + p(zz) \log p(zz). \end{aligned}$$

For the probabilities of trigrams in *English* texts see Pratt [149], for example, and the corresponding value of H_3 shall be given below.

An intuitive corroboration of what has been stated is provided by the situation in which an experiment, consisting of the draw of cards with three-letter combinations from $27^2 = 729$ urns, each of which contains cards with different trigrams starting from one and the same digram (equivalently, an experiment with an *English* book in which repeated efforts are made to select at random the digram coinciding with the last two letters chosen beforehand and to write down the letter appearing after it), leads to a 'sentence' such as the following:

IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID
PONDENOME OF DEMONSTURES OF THE REPTAGIN IS
REGOACTIONA OF CRE

This sentence represents 'third-order approximation' to *English*. It is also closer to the *English* language than its predecessors; it contains eight 'genuine' *English*

words and several easily pronounceable *English* sounding words (e.g., 'DEMONSTURES'). In analogy to this, we can also determine the entropy

$$\begin{aligned} H_4 &= H_{\alpha_1 \alpha_2 \alpha_3}(\alpha_4) = H(\alpha_1 \alpha_2 \alpha_3 \alpha_4) - H(\alpha_1 \alpha_2 \alpha_3) \\ &= -p(----) \log p(----) - p(---a) \log p(---a) \\ &\quad - \dots - p(zzzz) \log p(zzzz) + p(---) \log p(---) \\ &\quad + p(--a) \log p(--a) + \dots + p(zzz) \log p(zzz) \end{aligned}$$

that corresponds to an experiment to determine the next letter of an *English* text, provided that the *three* preceding letters are known. Corresponding to this quantity, an experiment that consists of drawing cards from 27^3 urns with four-letter combinations (or, an experiment with an *English* book similar to the one described above) leads to a 'sentence', which would contain mostly genuine *English* or *English*-like words. A still better approximation to the entropy of letters of an intelligible *English* text is given by the quantities

$$H_N = H_{\alpha_1 \alpha_2 \dots \alpha_{N-1}}(\alpha_N) = H(\alpha_1 \alpha_2 \dots \alpha_N) - H(\alpha_1 \alpha_2 \dots \alpha_{N-1})$$

when $N = 5, 6, \dots$. It is easy to see that with the growth of N the entropy H_N can only decrease (see p. 91). If it is further noted that all of H_N are positive, then from this it can be deduced that the quantities $H_{\alpha_1 \alpha_2 \dots \alpha_{N-1}}(\alpha_N) = H_N$ tend to a definite limit H_∞ as $N \rightarrow \infty$. This limit coincides with the limit H_∞ described in the preceding section (see p. 162).†

†The equality of the limit

$$\lim_{N \rightarrow \infty} \frac{H^{(N)}}{N} = \lim_{N \rightarrow \infty} \frac{H(\alpha_1) + H_{\alpha_1}(\alpha_2) + \dots + H_{\alpha_1 \dots \alpha_{N-1}}(\alpha_N)}{N}$$

considered in Section 4.2 to the quantity H_∞ introduced here follows from the fact that for large N almost all terms in the numerator of the fraction $H^{(N)}/N$ are close to

$$H_\infty = \lim_{N \rightarrow \infty} H_{\alpha_1 \alpha_2 \dots \alpha_{N-1}}(\alpha_N);$$

the only exception is the first few terms whose contribution to the total sum for N quite large is insignificant.

Thus, the sequence of 'specific entropies' $h_N = H^{(N)}/N$ as well as that of 'conditional entropies'

$$H_N = H_{\alpha_1 \alpha_2 \dots \alpha_{N-1}}(\alpha_N)$$

converge to one and the same limit H_∞ as $N \rightarrow \infty$. Also, $h_1 = H_1 = H(\alpha_1)$, but $H_N < h_N$ when $N > 1$ (since h_N equals the arithmetic average of N numbers, only the last of which is equal to but the rest are greater than H_N). Hence the quantities H_N , $N = 1, 2, 3, \dots$, approach the limit value H_∞ appreciably more rapidly than the quantities h_N (cf. footnote on p. 239).

From the results of Section 4.2 it follows that the *average number of elementary signals required for the transmission of one letter of an English text cannot be less than $H_\infty/\log m$* ; on the other hand, *a coding is possible for which this average number is arbitrarily close to the quantity $H_\infty/\log m$* (see p. 162). The difference $R = 1 - (H_\infty/H_0)$, expressing how much less than unity is the ratio of the 'limit entropy' H_∞ to the quantity $H_0 = \log n$, the latter characterizing the greatest amount of information that can be contained in one letter of an alphabet with a given number of letters, was designated by Shannon as the *redundancy* of a language (*English* in the case under consideration). The data, of which we shall speak below, compel us to assume that the redundancy of the *English* language (as also that of other *European* languages) appreciably exceeds 50%. Without claiming precision, we can say that the choice of the succeeding letter of an intelligible text is determined in more than 50% cases by the very structure of the language and, consequently, randomness is involved only to a comparatively small extent. It is specifically the redundancy of a language that enables us to contract the telegraphic language by discarding some words (articles, prepositions and conjunctions) that are easy to guess; it also allows us to reconstruct easily the true text even in the presence of a considerable number of errors in a telegram or misprints in a book.

In order to make clear the meaning of the quantity R , assume that an *English* text is encoded with the aid of a 27-ary code in which the same *English* letters are elementary signals. Such a 'code' is a certain method of shorthand writing of an *English* sentence by means of ordinary letters. In the case of a most efficient coding for writing an M -letter message we require on the average

$$\frac{H_\infty}{H_0} M = (1 - R)M$$

elementary signals (letters), i.e., in comparison to an ordinary written text we are able to economise by RM letters. This conclusion obviously does not imply that we can arbitrarily discard RM letters and then the remaining $(1 - R)M$ letters would suffice to reproduce the original message without error. In fact, for contracting a message by RM letters it is necessary to use a special 'very best' coding method, on applying which all letters of a message become independent and equally probable. Hence, it is clear that a text encoded here shall have the same character as the 'sentence' on p. 178, i.e., it will seem to be completely meaningless; it will be much more difficult to 'read' such a text than the 'sentence' given in the footnote on p. 180 (since the code words now correspond not to individual letters but directly to very lengthy 'blocks' of letters). We further note that in such a coding any error will be 'fatal': when decoded it will give us a new meaningful text and either we do not notice this, or even if we do so, we cannot make out what was actually written. As regards contracting a text by means of direct omission of a part of the letters that are chosen at random,

we can say in advance only that when *more than RM letters* are rejected we *cannot* a fortiori reproduce the original text without error. Specific experiments on the reconstruction of missing letters of *English* text have shown that usually a faultless reproduction is effected only if the number of discarded letters does not exceed 25% of their total number.

In particular for the *English* language we have redundancy estimates that are better than those for any other language. But even these estimates do not provide any especially reliable data. Clearly, the problem of redundancy estimation is equivalent to the problem of estimating the value H_∞ . But then how to determine the latter quantity? Using digram and trigram probabilities of the *English* language due to Pratt [149], Shannon [159] calculated the values of H_2 and H_3 . But it is clear that even H_3 is much far away from H_∞ . To obtain further estimates Shannon utilized the fact that different *English* words have also different probabilities of occurrence in a meaningful *English* text. The *English* word probabilities (estimated by frequency counts in a sufficiently lengthy sample of 'typical English text') are given in special frequency dictionaries of the *English* language (see, e.g., Eldridge [92], Dewey [90] or Thorndike [167]; cf., also [70]). The data in various frequency dictionaries are in satisfactory agreement. They show, for example, that *THE* is the most frequently used *English* word (its probability is close to 0.071); the next most probable word is *OF*, followed by *AND*, *TO* and so on. It is a remarkable fact that the probability p_n of the appearance of the n th word (in the decreasing order of word probabilities) is close to $0.1/n$ for quite a large number (a few thousands, in fact) of most probable words (this result shall be considered below in greater detail).

Using Dewey's frequency dictionary, Shannon [21] constructed an example of the so-called 'first-order word approximation' to the *English* language, i.e., of a sequence of genuine *English* words in which words are selected independently but with true probabilities of their appearance in the *English* text. This example is given below :

REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME
CAN DIFFERENT NATURAL HERE HE THE A IN CAME THE
TO OF TO EXPERT GRAY COME TO FURNISHES THE LINE
MESSAGE HAD BE THESE

It is clear that this sequence of *English* words represents a completely senseless *English* text.

Making use of more comprehensive data on the statistical characteristics of the written *English* language, Shannon constructed also a 'second-order word approximation' in which not only every word is selected in accordance with its probability to appear after a given preceding word but the statistical relationship between the two adjoining words is also taken into account (compare with the 'second-order letter approximation' which is related to entropy H_2 and is

described on p. 182). This new approximation has the form :

THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH
WRITER THAT THE CHARACTER OF THIS POINT IS
THEREFORE ANOTHER METHOD FOR THE
LETTERS THAT THE TIME OF WHO
EVER TOLD THE PROBLEM FOR
AN UNEXPECTED

Here also the whole text is senseless but various parts of it, composed of several adjoining words, compare favourably with the passages from the sensible *English* writing.

Let us now discuss the use of the term statistics for the approximate estimation of the entropies H_N of *English* language. It is clear that knowing the frequencies (probabilities) p_1, p_2, \dots, p_K of individual words (here K is the total number of words encountered in the text under consideration), we can calculate the 'first-order word entropy' by

$$H_1^{(\text{word})} = -p_1 \log p_1 - p_2 \log p_2 - \dots - p_K \log p_K.$$

Dividing the obtained value of $H_1^{(\text{word})}$ by the average number w of letters in an *English* word, we get an estimate for the conditional entropy H_w of order w . Expressly, it is easy to comprehend that $H_1^{(\text{word})}/w < H_w$ because the correlation among w letters of one word is appreciably stronger than that among w arbitrary sequences of letters in a meaningful text. On the other hand, the ratio $H_1^{(\text{word})}/w$ is certainly larger than the average information $H = H_\infty$ contained in one text letter, for the quantity $H_1^{(\text{word})}$ does not at all take note of the dependence existing among words (see p. 207 et seq.).†

According to Pratt [149], $w \approx 4.5$ for the *English* language. This enabled Shannon [159] to consider that $H_1^{(\text{word})}/w$ can be used as an approximate estimate of the entropy H_5 or H_6 . Shannon [159] also tried to extrapolate still further the series of values of H_N obtained. By this method he obtained a rather crude estimate of H_8 which agrees well also with some deductions from the existing cryptographic data. His results are summarized in the accompanying table (the above-mentioned values of H_0 and H_1 are also included here).

TABLE

H_0	H_1	H_2	H_3	H_5 or H_6	H_8
4.75	4.03	3.32	3.10	≈ 2.1	≈ 1.9

†Cf. also Urbach [169], in which Shannon's method is reconsidered and some estimates of entropies H_N other than those in [159] are derived. (In [169] the space between words is not included in the number of letters. This fact is, however, quite simple to take note of: see p. 203 et seq.)

Hence, it can be concluded that for the *English* language the redundancy R is in every case not less than $1 - (1.9/4.75) \approx 0.6$, i.e., it certainly exceeds 60%.

For a more precise estimate of the quantity R , it is further necessary to determine how much the quantity H_8 —the average information contained in a letter of a text given that the preceding seven letters are already known—differs from the limit value H_∞ . In other words, the problem in which we are interested is to what extent the arbitrariness in the choice of the next letter of an *English* text is essentially restricted by the knowledge of that part of the preceding text which is separated from this letter by more than seven letters (given that the following seven letters are also known). Since the average word length in the *English* language is close to four or five letters, i.e., it is appreciably less than seven letters, so the question here can only be the influence of statistical laws related to the dependence between successive *words of English* text (or even more general laws related to the succession of sentences). A direct solution of the problem we are interested in, through calculating the quantities H_9, H_{10}, \dots by using the formula given on p. 184, is impossible since for the determination of H_9 we need to know the probabilities of all nine-letter combinations, whose number is expressed by a 13-digit number (trillions!). Hence to evaluate the quantity H_N for large values of N , we have to confine ourselves to indirect methods. Here we shall briefly sketch a clever method of this sort, due to Shannon [159].

The 'conditional entropy' H_N is a measure of the uncertainty of experiment α_N , consisting of finding the N th text letter, given that the preceding $N - 1$ letters are known. This quantity naturally determines qualitatively the difficulty in guessing the N th letter when the preceding $N - 1$ letters are known. But the experiment for guessing the N th letter can be easily set up: for this it suffices to choose an $(N - 1)$ -letter fragment of a genuine *English* text and ask some one to guess the next letter.† The experiment can be similarly repeated many times; the labour involved here in finding the N th letter can be well estimated by means of the average value Q_N of the number of attempts entailed in the determination of the correct answer. It is clear that the quantities Q_N defined for different values of N are definite characteristics of the statistical structure of a language, in particular of its redundancy. In fact, in the case of zero redundancy, knowledge of an arbitrarily long fragment of text does not increase the probability of correctly guessing the next letter (this probability in all cases is $1/n$, where n is the number of alphabet letters). On the other hand, the equality of the redundancy to the quantity $1/m$ can be described quite roughly as a statement that every m th text letter is 'superfluous', uniquely reconstructed by the preceding $m - 1$ letters.

Obviously, the average number of efforts Q_N with the increase of N can only

†Shannon suggests that questions be put to a number of persons, and then the person who gives best answers, on the average, be selected, for it is considered here that the pursuit is carried out in a *most rational way*, i.e., with a complete knowledge of the statistical structure of a language.

decrease, the stoppage of this decrease will show the fact that the corresponding experiments have the same degree of uncertainty as experiments for greater value of N , i.e., that the 'conditional entropy' H_N has practically attained the limit value H_∞ . Starting from these arguments, Shannon set forth a number of similar experiments, in which N takes the values 1, 2, 3, . . . , 14, 15 and 100. In this connection, he observed that to find the 100th letter with respect to the 99 preceding ones is a considerably simpler problem than to find the 15th letter with respect to the 14 preceding ones. Hence it can be concluded that H_{15} is substantially larger than H_{100} , i.e., that it is by no means possible to identify H_{15} with the limit value H_∞ . The same experiments were later conducted on a somewhat larger scale by Burton and Licklider [81] for $N = 1, 2, 4, 8, 16, 32, 64, 128$ and $N \approx 10,000$. From their data it is possible to infer that the quantity H_{32} (and, of course, also H_{64} and H_{128}) practically does not differ from $H_{10,000}$, while the 'conditional entropy' H_{16} is still appreciably larger than this quantity. Thus, it can be assumed that with increasing N the quantity H_N decreases up to values of N of order 30, but with further growth of N it remains practically invariant; hence, instead of the 'limiting entropy' H_∞ we can speak, for instance, of the conditional entropy H_{30} or H_{40} .

The experiments on guessing letters not only enable us to predetermine the comparative magnitudes of the conditional entropies H_N for distinct N , but also provide an opportunity to estimate even the values of H_N . This opportunity is connected with the fact that by the data of such experiments it is possible not only to determine the average number Q_N of trials required to guess the N th text letter with respect to prefixes $N - 1$, but also to estimate the probabilities (i.e., limiting frequencies) $q_N^1, q_N^2, \dots, q_N^n$ of guessing correctly a letter by the 1st, 2nd, 3rd, . . . , n th trials (where $N = 27$ is the number of alphabet letters). It is obvious that $Q_N = q_N^1 \times 1 + q_N^2 \times 2 + \dots + q_N^n \times n$. It is also easy to understand that the probabilities $q_1^1, q_1^2, \dots, q_1^n$ are the probabilities $p(a_1), p(a_2), \dots, p(a_n)$ of alphabet letters a_1, a_2, \dots, a_n arranged in *order of decreasing probabilities*. In fact, if no letter preceding the letter x to be guessed is known to us, then it is natural to assume first that x coincides with the most widely used letter a_1 (the probability of guessing correctly here being $p(a_1)$); next we must assume that x coincides with a_2 (the probability of correct guessing here being $p(a_2)$), and so on. This implies that the entropy H_1 equals the sum

$$-q_1^1 \log q_1^1 - q_1^2 \log q_1^2 - \dots - q_1^n \log q_1^n.$$

If, however, $N > 1$, then it can be shown straightaway that the conditional entropy H_N does not exceed the sum

$$-q_N^1 \log q_N^1 - q_N^2 \log q_N^2 - \dots - q_N^n \log q_N^n. \quad (*)$$

The inequality (*) follows from the fact that the quantities $q_N^1, q_N^2, \dots, q_N^n$ result from some averaging of the probabilities of the outcomes of experiment α_N (see Shannon [159] or Savchuk [155]). On the other hand, somewhat deeper (but at the same time not tedious) reasonings enable us to show that the sum

$$(q_N^1 - q_N^2) \log 1 + 2(q_N^2 - q_N^3) \log 2 \\ + \dots + (n-1)(q_N^{n-1} - q_N^n) \log (n-1) + nq_N^n \log n \quad (**)$$

for every N is not greater than the conditional entropy H_N .† Thus, the expressions (*) and (**) (made up of the probabilities $q_N^1, q_N^2, \dots, q_N^n$, which can be estimated by the data on guessing experiments) define the bounds between which H_N must be contained.

It is also necessary to keep in mind that both the estimates (*) and (**) are obtained with the assumption that $q_N^1, q_N^2, \dots, q_N^n$ are those probabilities of guessing a letter with respect to $N-1$ preceding letters in first, second, third, ..., trials, which prevail in the presumption that guessing always identifies the next letter *most appropriately*—with full regard to all statistical laws of the given language (see footnote on p. 188). In the case of real experiments, however, any mistake in the strategy of guessing (i.e., the variance of a letter identified by it from the required one, which stems from the exact statistics of language) inevitably leads to an overstatement of both the sums (*) and (**). Hence, it is specifically expedient to take note of only the data of the ‘most successful guesser’, since this overstatement will be the least for him. However, since every guesser deviates sometimes from the best guessing strategy, it is practically impossible to consider (**) as a completely reliable lower bound on true entropy (in distinction to the upper bound (*), which because of erroneous guesses may only become still larger).

Furthermore, the values of (*) and (**) unfortunately do not come closer together indefinitely with increasing N (starting with $N \approx 30$ these sums in general cease to depend on N); hence the estimates of redundancy for a language obtained here are rather loose.†† In particular, Shannon’s experiments [159] show only that H_{100} is apparently contained between 0.6 and 1.3 bits. Hence, it can be concluded that the redundancy

$$R = 1 - \frac{H_\infty}{H_0} \approx 1 - \frac{H_{100}}{\log 27}$$

†The derivation of this result due to Shannon has been further elucidated by Savchuk [155] and Maixner [124].

††See Savchuk [155], where the completely artificial ‘languages’ are constructed, for which Shannon’s entropy estimate (*), or correspondingly (**), is exact.

for *English* is almost certainly higher than 70% and quite probably may be close to 80% or even higher. The experiments due to Burton and Licklider [81] led to similar results: according to their data, the true value of redundancy for *English* lies somewhere between $\frac{2}{3}$ (i.e., 67%) and $\frac{4}{5}$ (i.e., 80%). Finally, Piotrovskii, Bektaev and Piotrovskaya [148] indicate the following results of the letter-guessing experiments for *English* language: $72\% \leq R \leq 84\%$. They have given also specific results for three different types of *English* text; these results will be discussed later.

Shannon's method of entropy estimation by guessing experiments was considerably improved subsequently by Kolmogorov and by Cover and King. This interesting development heralding the information theory approach to the written language will be discussed below in this section. At this stage, we shall only remark that the estimate of entropy H_∞ of the *English* text due to Cover and King [86], which is apparently the best available at present (but is also only preliminary), shows that H_∞ is smaller than 1.3 bits. It agrees well with Shannon's estimate and also shows that the redundancy R of *English* is not lower than 70%.

Before we undertake an examination of the results for various other languages, it is appropriate to make a few additional comments. A mention has already been made of Shannon's recommendation to take note of only the results of the subject who guesses the letters most successfully. It is clear that the amount of success in guessing characterizes the degree of guesser's understanding (usually intuitive) of the statistical laws of language, i.e., a 'feel of language' intrinsic to a given subject (or 'feel of style' of a given author, whose text is used for letter-guessing; cf. the remark in Kolmogorov [15] to the effect of one of the guessers, who has obviously a particularly developed literary flair, having a 'telepathic relationship with the author'). Hence from Shannon's view point, the differences between the results of different subjects, participating in letter-guessing experiments, have to be regarded as undesirable (though, unfortunately, these are unavoidable), because these experiments rely on an 'ideal guesser' having the maximum amount of familiarity with the intrinsic statistics of the given language. It was, however, observed by Attneave [42] that in fact the differences in the entropy values obtained by different subjects in letter-guessing experiments are of definite interest, because such differences characterize quantitatively the level of language fluency, vocabulary and factual knowledge possessed by different subjects. In fact, efforts have already been made to utilize the results of letter-guessing experiments for an objective measurement of the extent of one's grasp over a foreign language ([161]; see also [112]) or one's mother tongue (see, e.g., [135] which describes experiments on letter guessing of a highly specialized text by a few groups of persons with a highly diverse level of practice in reading a text of similar contents). Weltner [173] adduces quite rich material related to similar evaluations of 'subjective information' contained in a given text (for a given person), and emphasises the great value of such subjective information for educational purposes.

Weltner started from a slightly modified version of Shannon's letter guessing method† and used it for the determination of subjective information contained in diverse types of texts (a scrambled text, scientific paper, poem, fiction prose text, newspaper text, ordinary and programmed textbooks) for different categories of readers (high school students from different schools, students of a teacher training college etc.). The letter-guessing experiments due to Nemetz and Simon [134], which will be described in detail later, were also carried out on a collection of different guessers (e.g. specialists in mathematical statistics, teen-aged high school students, teachers of literature and mathematics in high school, and so on). In the present book, however, we shall not dwell upon the study of subjective information, which is more predominantly a psychological than a purely mathematical notion.

Now let us bring into consideration the results related to various foreign languages. It is clear that for all languages that make use of the *English* alphabet, the maximum information H_0 that can be conveyed by one letter of a text (including space) has one and the same value:††

$$H_0 = \log 27 \approx 4.75 \text{ bits.}$$

However, the frequencies of the appearance of various letters and many hyphen letter combinations are obviously different in different languages. Thus, for example, by arranging all letters in order of increasing probabilities (starting with the most frequent of them), we arrive at a sequence of letters beginning with—*ETAONRI* . . . in the case of *English* language, where '—' denotes the space between words (see p. 179 above). Moreover, in the case of the *German* language the corresponding sequence will begin with—*ENISTRAD* . . . , and in the case of French with—*ESANITUR* . . . (see [75]).

The average word-length defining the probability of 'space' is appreciably greater in the *German* language than in the *English* or the *French*; the letters *W* and *K* are encountered comparatively frequently in the *German* and the *English*

†The main difference between Shannon's and Weltner's experiments is related to the fact that Weltner considers a 32-letter alphabet (including also some punctuation marks) and reduces the guessing process to 'binary choices'.

††There are, however, languages which use 'English-like' (or, more correctly, 'Latin-like' alphabets but of a different number of letters. Several *European* languages do not use all the *English* letters (e.g., two letters *K* and *W* are not used in *Spanish* and five letters *J*, *K*, *W*, *X* and *Y* are not used in *Italian*). On the other hand, there are *European* alphabets which include some supplementary letters that differ from ordinary *English* letters by special marks (e.g., letters \ddot{a} and \ddot{o} are used in *Finnish*, *Swedish*, and *German*, \grave{e} , \acute{e} and $\text{\textit{ç}}$ in French, \tilde{n} in Spanish, \emptyset in *Norwegian*, $\text{\textit{å}}$ in *Swedish* and *Norwegian* and so on). We may, however, agree to include into *Spanish*, *Italian* and related alphabets all the missing *English* letters as letters having zero probability (in fact, these letters may occur occasionally in the corresponding texts when foreign names are mentioned). We may also agree to make no distinction between supplementary letters and related *English* letters (and in the case of *German* to write \ddot{a} , \ddot{o} and \ddot{u} as *ae*, *oe* and *ue*). With this approach our statement regarding H_0 shall remain correct for all languages that use the *Latin* alphabet.

languages, but have very low probability in the *French*; the combination *TH* is quite prevalent in the *English* language and so is *SCH* in the *German* language, but in other languages both these combinations are considerably less frequent; the letter *C* is almost always followed either by the letter *H* or by *K* in the *German* language, but not in the *English* or *French* and so on. Therefore, the first-order, second-order, . . . , letter approximations, whose typical examples for the *English* have been set out on pp. 181, 182 and 183, will have quite different forms for different languages (though, of course, the zero-order approximation remains the same for all languages). Abramson [1] has presented the approximations of the first three orders for *French*, *German* and *Spanish*; the corresponding third-order approximations (related to entropy H_3) are given by the following examples:

JOU MOUPLAS DE MONNERNAISSAINS DEME US VREH BRE TU
DE TOUCHEUR DIMMERE LLES MAR ELAME RE A VER IL
DOUVENTS SO (in the French)

BET EREINER SOMMEIT SINACH GAN TURHATTER AUM WIE
BEST ALLIENDER TAUSSICHELE LAUFURCHT ER
BLEINDESEIT UBER KONN (in the German)

RAMA DE LLA EL GUIA IMO SUS CONDIAS SU E UN-
CONDADADO DEA MARE TO BUERBALIA NUE
Y HERARSIN DE SE SUS SUPAROCEDA (in the Spanish)

The three passages are senseless in any language, but nevertheless any one who has even a rudimentary knowledge of the indicated languages can easily determine to which language each of these passages approximates.

Using the letter frequency tables for different languages it is possible to compute the corresponding values (in bits) of the entropy H_1 . Some of the results are listed in the accompanying table.

	Language					
	<i>English</i>	<i>German</i>	<i>French</i>	<i>Spanish</i>	<i>Italian</i>	<i>Portuguese</i>
H_1	4.03	4.10	3.96	3.98	3.90	3.91

(see Barnard [73] and Manfrino [128]).† In all the cases the value of H_1 is seen to be appreciably less than $H_0 = \log 27 \approx 4.75$ bits, and the values of H_1 for

†The values of H_1 for the *Italian* and *Portuguese* languages were computed by Manfrino for three different types of texts and for the alphabet which did not include the space. For our purpose, we have taken the mean of the three Manfrino's values of H_1 and then recalculated the values of H_1 for the alphabet which includes space. The equation that enables us to make such a recalculation shall be given below in this section (see p. 204).

different languages do not strongly differ here from each other. Of the examples cited, H_1 has the highest and least values, respectively, for the *German* and *Italian* languages. The higher value for the *German* is apparently due to the fact that the average word-length has the largest value in this language and hence the probability of space (which is the most frequent letter in all languages) is smaller in *German* than in all other considered languages. As regards the *Italian* language, there are five different *Latin* letters which have a zero probability in *Italian*; hence, the number of outcomes of the letter-guessing experiment α_1 is here smaller than in the other languages. The values of H_1 for some other languages not included in the above table can be found in [128], [113] and other references listed at the end of the book.

To compute the entropies H_2 and H_3 for various languages it is necessary to know the corresponding digram and trigram probabilities, i.e., the probabilities of two-letter and three-letter combinations. Some data on these probabilities for a number of languages may be found in Pratt [149]. Further data have been published recently by several authors engaged in the specific task of calculating the entropies H_2 and H_3 . For example, Petrova [143] utilized a sample made up from miscellaneous *French* texts, consisting of 30,000 characters, for the evaluation of the French digram and trigram probabilities. With a similar objective, Manfrino [128] used three 10,000-letter samples from the *Italian* scientific, history and newspaper texts as well as three samples of about the same length from the *Portuguese (Brazilian)* scientific, fiction and newspaper texts; however, in contrast to Petrova, he did not include the space in the number of alphabet letters. Lebedev and Garmash [120] treated a passage from L. N. Tolstoy's novel *War and Peace*, containing roughly 30,000 letters; they catered for space in the number of letters and considered the *Russian* alphabet as consisting of 32 letters (they made no distinction between the letters e and \ddot{e} , \mathcal{B} , and \mathfrak{z} , which is also the practice followed in almost all *Russian* telegraph codes). Wanas *et al.* [128] analyzed a sample from one of the *Arabic* newspapers made up of 64,000 letters (the Arabic alphabet selected for this study consisted of 32 letters). There are also more examples of the related studies which we shall not mention here and refer the reader to the 'references' at the end of the book. The results of investigations cited are listed in the following table which contains, for the sake of completeness, the values of H_0 , H_1 , H_2 , H_3 and the entropy values for the *English* language as well:

	Language					
	<i>English</i>	<i>French</i>	<i>Italian</i>	<i>Portuguese</i>	<i>Russian</i>	<i>Arabic</i>
H_0	4.75	4.75	4.39	4.52	5.00	5.00
H_1	4.03	3.95	3.90	3.91	4.35	4.21
H_2	3.32	3.17	3.32	3.35	3.52	3.77
H_3	3.10	2.83	2.76	3.20	3.01	2.49

The different columns of the table do not differ sharply from each other, which does not seem to be surprising. However, the table can be hardly used for evaluating the redundancy for the written text in the indicated languages, since H_3 is obviously still quite far from the limiting value H_∞ .

An approximate estimate of H_N with $N > 3$ can be obtained with the aid of Shannon's method (based on the letter-guessing experiments), or some modified variant of it. A number of attempts undertaken from this motivation are described in the existing literature. However, the fact remains that most of the results achieved in this direction are of an even more preliminary character than the rough results for the *English* language presented above.

The redundancy for *German* has been investigated quite thoroughly by Küpfmüller [118]. By using the available data on the frequencies of occurrence of different syllables and words in *German* and performing some experiments on guessing the succeeding syllables or words of a *German* text with respect to the known preceding excerpt, Küpfmüller inferred that for the *German* language $H_\infty \approx 1.3$ bit. This implies that the redundancy R of this language is close to

$$1 - \frac{1.3}{4.75} \approx 0.73,$$

a value having the same order of magnitude as the estimate of the redundancy for *English* deduced above. The value of H_2 for *German* may be found, in particular, in [93]. The results of letter-guessing experiments for *German* language are presented in Piotrovskii, Bektaev and Piotrovskaya [148]. These include two estimates of the redundancy R deduced from Shannon's upper and lower entropy limits (*) and (**) (see pp. 189 and 190) related to three different types of the *German* text (conversational language, fiction and various business texts). The average results of Piotrovskii *et al.* for the *German* language at large are very close to the corresponding results for the *English* language: they indicate that $71\% \leq R \leq 85\%$.

A study of the entropy and redundancy of the *French* language has been made in great depth by Petrova [143]. Her results related to the values of the entropies H_N for $N = 1, 2$ and 3 have been briefly described above. To determine the values H_N , when N is large, the letter-guessing experiments were employed, applying partly a refinement of the procedure suggested by Kolmogorov, of which we shall say more later on. The deductions in [143] have yielded the estimate $H_\infty \approx 1.40$ bits and, consequently, $R \approx 71\%$. A similar (but somewhat cruder) study of the redundancy for the *Swedish* language has been carried out by Hansson [128] leading to the result that $H_\infty \leq 2$ bits and $R \geq 1 - (2/\log 30) \approx 59\%$ (Hansson considered the 29-letter *Swedish* alphabet, i.e., the total number of letters accounted for by him, with the inclusion of word space, came to 30). For several other evaluations of the entropies and redundancies of various languages the reader is referred to [113], [128], [147] and [148]. In [148], in the particular,

there are adduced the estimates of redundancy R from above and below for three diverse types of texts (colloquial, fictional and scientific) written in seven languages (*English, German, Russian, French, Polish, Rumanian and Kazakhian*). These estimates have been obtained with the aid of letter-gussing experiments on the basis of relations (*) and (**) and deviate only slightly from each other irrespective of the seven languages involved.

The results indicated show that the redundancy estimates of most of the *European* languages do not diverge widely from each other but this fact does not permit us to conclude that the same must hold also for the languages which are either quite apart in their linguistic structure or differ sharply in their alphabets. In this connection, the investigation of Newman and Waugh [138] is of interest. They have endeavoured to compare the entropies H_N and redundancies R for three languages with appreciably distinct numbers of alphabet letters: for the Polynesian *Samoyan* language, whose alphabet contains altogether 16 letters (nearly 60% of which are vowels), for the *English* and the *Russian*. In the case of *Russian*, the specially chosen texts were printed in old orthography (used in Russia up to 1918), using a 35-letter alphabet. It is natural for the quantity H_0 to have highly different values for these three languages (see the accompanying table). The values of H_1 listed in this table for the three languages differ still

	<i>Samoyan</i>	<i>English</i>	<i>Russian (old orthography)</i>
H_0	$\log 17 \approx 4.09$	$\log 27 \approx 4.75$	$\log 36 \approx 5.17$
H_1	3.40	4.08	4.55
H_2	2.68	3.23	3.44

more sharply. (The letter frequencies used in the evaluation of H_1 have been compiled by Newman and Waugh on the basis of an analysis of the same passage from three translations of the *Bible* having the length of nearly 10,000 characters.) The variations in the values of H_1 roughly signify that the probability distribution of individual letters is most uniform in *Russian*, but in *Samoyan* it is most nonuniform. To a considerable extent this conclusion is explained by the fact that in *Samoyan* the average word-length is quite small: it is close to 3.2 letters against 4.1 letters for *English* and 5.3 letters for *Russian* texts considered. Hence, a word space, the most frequent character, has the largest probability in *Samoyan*, less in *English*, and still less in *Russian*. However, the values of H_2 for the three languages are found to be closer than those of H_1 : the two-letter correlations in *Russian* are more stringent than in *English* and still more so than in *Samoyan*.

Unfortunately, the successive values of H_N for $N > 2$ given by Newman and Waugh are not reliable, for these have been obtained by them by means of a disputable method developed by Newman and Gerstman [137]. However, their conclusions concerning the comparative values of H_N for the three languages strike to be plausible. According to these conclusions, the values of H_N decrease

most rapidly in *Russian* and most slowly in *Samoyan*; as a result, starting from approximately $N = 10$ the values of H_N (and, consequently, also of H_∞) for the three languages are found to be sufficiently close to each other. This signifies that the average amount of information per text letter for three languages having appreciably distinct numbers of alphabet letters is approximately the same. If this conclusion is true, then it obviously implies that the redundancy is considerably greater for languages affluent in the number of distinct letters than for those with meagre alphabets.

Note also that in all *European* languages the vowels are considerably more frequent than the consonants. This fact is responsible for significant differences in the frequencies of individual letters, which appreciably affect the value of the 'first-order entropy' H_1 (and also the 'limit entropy' $H = H_\infty$ and the redundancy R) of a language. The position is different in a number of *Oriental* languages. For instance, in *Hebrew* the vowels are not used at all: they are omitted in the written text and are supplied by the reader 'according to sense' (this is plausible by virtue of the redundancy of a language). It is clear that the statistical structure of a text written in this language differs sharply from that encountered in *European* languages, in view of which the values of all the information-theoretic characteristics of a language may take here quite different values (in particular, the redundancy must reduce appreciably). As an illustration of this remark, a reference may be made to Bluhme [78], who compared statistical characteristics of a collection of three-letter words from *Hebrew* and *English* and discovered that for this collection

$$H_3^{(\text{Heb})} \approx 3.73 \text{ (bit/letter)} \text{ and } R_3^{(\text{Heb})} \approx 1 - \frac{H_3}{H_0} \approx 0.16,$$

whereas

$$H_3^{(\text{Eng})} \approx 0.83 \text{ (bit/letter)} \text{ and } R_3^{(\text{Eng})} \approx 0.82.$$

The entropy of individual *Indian* languages was also studied in detail in the sixties, in the first place the *Dravidian* languages prevalent in South India and belonging to the stock of the most ancient human languages. In [160], starting from the statistical language data (and taking note of the correction introduced in [74]), the values of lower order entropies are found for several *Indian* languages and Shannon's 'method of guessing experiments' is also used to estimate the values of H_N when N is comparatively large. In this connection, we note that in comparison to works related to *European* languages, new difficulties arose because of some uncertainty of alphabets in a majority of the considered languages. Thus, for example, in *Tamil* there exist both classic and modern alphabets; in the modern alphabet (close to the alphabets of a number of other *Indian* languages) there exist 12 vowels, 18 consonants, 216 unified consonant-vowels

and one more unpronounced symbol (Aitham) for a special purpose. In Siro-money's work [160] Aitham is completely ignored, and 'consonant-vowels' are considered as pairs of letters; however, such an approach to the *Tamil* language is not the only one possible. We shall set forth later (see p. 214) some of the results of studies devoted to *Indian* languages.

Finally, let us note that the differences in the presently available estimates of the values of the entropy $H = H_\infty$ (or the quantities H_N , where N is moderately large) manifested for different *European* languages by means of the 'guessing experiment method' are, as a rule, appreciably smaller than the accuracy of the respective estimates determined by the difference between the lower and upper bound expressions (*) and (**) for the N th order entropy.

Thus, the Shannon method turns out to be clearly inadequate for determining differences in the specific entropy (per letter) for different languages, although the existence of differences in the average word-length and the length of parallel texts, having the same content, for different languages (see, Ramakrishna and Subramanian [151], and also the last reference in [160])† creates an impression that these differences in specific entropies may be of an order of 10–20%. The same can also be said of the differences in the entropies of texts of different characters (in particular, due to different authors) written in the same language: it is quite obvious that the differences in these may be sufficiently large, but they may also be detected by means of the Shannon method only in the most exclusive cases (like those to which are related the works of Frick and Sumby or Fritz and Grier, mentioned on p. 212)

In this connection, it is highly desirable to have a more precise method for determining the entropy of a language. Kolmogorov stated that such a method is comparatively simple to obtain by further sharpening the 'guessing method'. In particular, Kolmogorov noted that in principle the guessing method (with the assumption that the guessing subject always follows an 'optimal strategy', which stems from a complete knowledge of all statistical regularities inherent in a given language) enables us to obtain not only an estimate of upper and lower bounds of entropy, but also an *exact estimate* of the value of this quantity. In fact, assume that while guessing one does not name only one alphabet letter each time selected in accordance with the order in which the probabilities of letter appearance decrease, but directly indicates all conditional probabilities $p_1^N, p_2^N, \dots, p_n^N$ of the occurrence of the 1st, 2nd, \dots , n th alphabet letter (given that $N - 1$ text letters preceding it are known).

†However, the two indicated works are in fact of interest only from the viewpoint of the formulation of problems, but not from the viewpoint of the specific results obtained here. The reason is that for evaluating the 'efficiency' of different languages, there has been employed only a comparison of 'first-order entropies' H_1 of these languages, without taking any account of the statistical relationship between various successive text letters, which is extremely important in linguistic structure.

Suppose now that this experiment is repeated many times and each time the value of the quantity $-\log p_k^N$ is calculated, where k is the number of the letter that actually appeared. Thus, in every individual experiment of 'guessing' from n given numbers p_1^N, \dots, p_n^N (where n is the number of alphabet letters), in fact only one number is taken into account, but expressly the one which is not known beforehand. It is now easy to show that if the conditional probabilities are always determined exactly, then the average value of the enumerated quantities $-\log p_k^N$ (i.e., the sum of all such quantities which are determined in a large number of M experiments, divided by M) for unboundedly increasing M tends to the true entropy H_N of one text letter.

This method seems to be completely impracticable: it is inconceivable to demand of guessing subject that every time he would determine the entire collection of conditional probabilities of all possible letters and, in addition, that none would be in error (cf., in this connection, the analysis of the work by Cover and King [86], given below). It is, however, essential that any error in the specified values of conditional probabilities cause only an *increase* in the corresponding sum of the values $-\log p_k^N$ (this statement, as it is easy to show, follows from the fact that (*) on p. 189 gives an upper bound of H_N). Hence it is completely permissible to restrict beforehand the set of probability distributions, which can be named in guessing, and with that substantially facilitate its performance; here the sum thus obtained of values $-\log p_k^N$, divided by the number M of experiments, is all the same the *upper bound* on the true entropy H_N .

In real experiments conducted under the guidance of Kolmogorov on *Russian* literary texts, the following forecasts were provided for guessing (cf. [154]) :

- (i) one specific (say, k th) alphabet letter would certainly be next letter;
- (ii) one of the two or three alphabet letters to be indicated in guessing would certainly be the next letter;
- (iii) one specific (say, k th) alphabet letter would probably (but not certainly) be the next letter;
- (iv) one of the two or three letters to be indicated in guessing would probably be the next letter;
- (v) moreover, the guessing would permit one to say that one does not know which will be the next letter.

It was also assumed that each of these statements is equivalent to the choice of the following conditional probability distribution for the succeeding text letter:

- (i) the k th letter has some preassigned large probability P ; however, for the i th letter, where $i \neq k$, the probability of its appearance is taken as $p'_i = p_i / [(1 - P)(1 - p_k)]$, where p_i and p_k are unconditional probabilities of the i th and k th alphabet letters (these probabilities are known for

- many languages, including the *Russian* language; for *English* these are listed in the table on p. 179);
- (ii) the two or three letters chosen have the same conditional probability $P/2$ or $P/3$. The remaining letters have, as before, the probabilities p'_i , proportional to their unconditional probabilities p_i ;
 - (iii) the k th letter has some fixed probability Q (smaller than P !), but the i th letter, for $i \neq k$, has the probability $p'_i = p_i \times [(1 - Q)/(1 - p_k)]$;
 - (iv) the two or three letters chosen have the same probability $Q/2$ or $Q/3$, but the remaining letters have probabilities proportional to their unconditional probabilities;
 - (v) the conditional probability of the appearance of i th alphabet letter for all i is taken to be equal to its unconditional probability p_i .

The probabilities P and Q remain so far undetermined; however, since any inaccuracy in the indicated conditional probability distribution may only increase the estimate obtained for H_N , it is completely admissible to select these two probabilities, according to the known experimental results, in such a way that the sum of all quantities $-\log p_k^N$ (where p_k^N is the predicted conditional probability of the letters having actually appeared) is the *least possible*.

It is easy to calculate that, with such definitions of the probabilities P and Q , the final estimate of the entropy H_N is given by the formula

$$H_N \approx \frac{1}{M} [M_1 h_1 + M_2 h_2 + M'_1 + M'_2 \log 3 + S],$$

where M is the total number of experiments; M_1 is the number of forecasts of type (i) or (ii); M_2 is the number of forecasts of type (iii) or (iv); M'_1 is the number of forecasts of type (ii) or (iv), in which the two possible letters are indicated; M'_2 is the number of forecasts of type (ii) or (iv), in which the three possible letters are indicated; $h_1 = -q_1 \log q_1 - (1 - q_1) \log (1 - q_1)$, where $q_1 = m_1/M_1$ and m_1 is the number of errors in the forecasts of types (i) and (ii); $h_2 = -q_2 \log q_2 - (1 - q_2) \log (1 - q_2)$, where $q_2 = m_2/M_2$ is the average fraction of errors in the forecasts of types (iii) and (iv); finally, S is the sum (extended over all cases of errors in the forecasts of types (i), (ii), (iii) and (iv), and all 'rejections', i.e., 'forecasts of type (v)) of the expressions $-\log p_i^*$, where p_i^* is either the 'unconditional probability' p_i of a letter having actually appeared (in the case of forecasts of type (v)), or is the 'forecasted probability' p'_i divided either by $1 - P$ (in the case of forecasts of types (i) and (ii)), or by $1 - Q$ (in the case of forecasts of types (iii) and (iv)).

The above equation appears, at the first glance, comparatively intricate, but in practice it is found to be sufficiently convenient and does not involve very cumbersome calculations. Guessing experiments of such kind were carried out in the statistical laboratory of Moscow State University, which enabled the

experimenters to obtain in the case of classical nineteenth century *Russian* prose of S. T. Aksakov (*The Childhood of Bagrov the Grandson*, a novel) and I. A. Goncharov (*Literary Party*, a short story) the bound on the specific entropy H_∞ (not differing, say, from H_{50}) of the order of 1—1.2 bit. This bound is apparently quite precise (probably exceeding the true value of H_∞ by not more than 10—15%). According to this, the value of redundancy for the literary language of *Russian* classical prose is close to 80%.

The recent work of Cover and King [86] (containing an extensive bibliography) is quite similar to Kolmogorov's investigation. They also used a refined variant of Shannon's letter-guessing experimental technique. The main idea underlying their work is, in fact, identical to Kolmogorov's postulate (see pp. 198-99) that if the guesser is asked to list all the conditional probabilities $p_1^N, p_2^N, \dots, p_n^N$ of the occurrence of various alphabet letters at the N th place after $N - 1$ known letters, then, in the case of an ideal error-free guessing, the average value of $-\log p_k^N$, where k is the number of the letter that has actually appeared, will give an exact estimate of the entropy H_N . (This result was obtained in a slightly different form by Cover and King independently of Kolmogorov.)

The procedure proposed by Cover and King has the form of the following 'gambling scheme'. Let us consider a subject having at the beginning the capital $S_0 = 1$ dollar. The subject knows $N - 1$ letters of the text and wants to place bets on the next letter. He is allowed to wager any percentage $p_i S$ of his capital S on the i th alphabet letter (where, of course, $p_1 + \dots + p_n = 1$ and $n = 27$ is the number of different *English* letters including space). If the i th letter appears as the N th letter of the text, the subject wins the capital $np_i S = 27p_i S$. The process is repeated many times; let us denote by S_M the subject's capital after M bets. If the subject permanently wagers the same capital $S/27$ on every letter, then he will preserve the same capital after all bets. If, however, he distributes his stakes inhomogeneously using the known statistics of the language, then his capital will increase with a very high probability. Cover and King showed that the optimal gambling strategy is to wager every time a percentage of the current capital in proportion to the conditional probability of the next symbol, i.e., to select p_1, \dots, p_n equal to the conditional probabilities of alphabet letters when preceding $N - 1$ letters are known. If the stakes are selected by following this strategy, then the capital S_M will increase with probability 1 and the quantity

$$\log n \left(1 - \frac{1}{M} \log_n S_M \right) = \log_2 27 \left(1 - \frac{1}{M} \log_{27} S_M \right)$$

will tend to H_N as $M \rightarrow \infty$. It is clear that $S_M = (27)^M p_{k_1}^N p_{k_2}^N \dots p_{k_M}^N$, where k_i is the number of the letter which actually appeared in the i th bet, and

$$\log_2 27 \left(1 - \frac{1}{M} \log_{27} S_M \right) = -\frac{1}{M} \left(\log p_{k_1}^N + \log p_{k_2}^N + \dots + \log p_{k_M}^N \right);$$

therefore, this expression due to Cover and King is evidently equivalent to Kolmogorov's proposition formulated on p. 199. If, however, p_i differ from true conditional probabilities of alphabet letters (which are not known exactly to any real subject), then the increase in the capital will be slower and hence the limit of the quantity indicated, as $M \rightarrow \infty$, will give an estimate of H_N from above.

Cover and King performed an actual experiment on evaluating the entropy of *English* language by the procedure described. The text used was taken from the same book, *Jefferson the Virginian* by Dumas Malone, which was employed in Shannon's letter-guessing experiments [159]. Twelve persons were selected and all of them were given the same part of the book from its beginning and up to an abrupt end in the middle of a word. They were allowed to read as much of the book as they desired up to the selected end in order to familiarize themselves with the style of author's writing. Each person was also allowed to use tables of *English* letters, digram and trigram probabilities (Cover and King noted, however, that the use of tables did not help to improve the results). Under these conditions, all 12 'gamblers' were asked to distribute stakes $p_1S, p_2S, \dots, p_{27}S$ on the possible appearances of the next letter (here $p_1 + p_2 + \dots + p_{27} = 1$ and S is the current capital of a gambler). After every bet the actual next letter of text was exposed, capital of every gambler was recomputed, and the whole procedure was repeated again. The game was finished after 75 bets (work at a computer terminal for any of 12 subjects took about 5 hours). Since the number $N - 1$ of the text letters known beforehand was quite large in this experiment, the estimate obtained here refers directly to H_∞ . The details of 'gamblers' decisions' were not set forth by Cover and King. However, it is clear that the subjects involved in experiments hardly proposed 27 different numbers p_1, p_2, \dots, p_{27} at every bet, but they apparently selected one probability distribution from a small set of simple model distributions similar to (i)–(v) explicitly formulated by Kolmogorov.

The results of all gamblers (the 'final capital' S_{75} and the resultant entropy estimate) are listed at the end of Cover and King's paper. All estimates range between 1.3 and 1.9 bits per letter. Moreover, the best subject estimate of H_N , the average capital estimate (based on the total capital of gamblers) and the so-called committee gambling estimate (based on a more complicated averaging of the results of different gamblers)—all lead to the value $H_\infty \approx 1.3$ bit/letter (i.e., lead to the inference that, for the written *English*, H_∞ is in fact smaller than 1.3 bit/letter, in other words, that $R \geq 73\%$). These results agree well with the results of similar Kolmogorov's investigations related to the *Russian* language.

Cover and King also attempted a similar experiment for a different type of text, namely for a text from the book, *Contact : First Four Minutes* by Leonard Zunin, which happened to be of greater professional interest to the selected 'gamblers' than the Dumas Malone book. This experiment was not concluded

till the publication of Cover and King's paper [86]. However, the results of first two subjects yielded a slightly lower entropy estimate. At present it is not possible to decide whether this small difference in the entropy estimate for the two books is real or fictitious.

The procedure suggested by Cover and King was applied recently by Nemetz and Simon [134] for estimating the entropy of the *Hungarian* language. The *Hungarian* alphabet is different from the *English* alphabet: the former includes 9 additional 'accented' letters (\acute{a} , \acute{e} , \acute{i} , \acute{o} , \acute{o} , \acute{u} , \acute{u} , and \acute{u}), but excludes the letters q , w , x and y which may appear only in foreign words and foreign names. However, such foreign words and names are becoming more frequent in the modern *Hungarian* texts; therefore, Nemetz and Simon did not exclude these letters from the *Hungarian* alphabet. On the other hand, they identified the letters \acute{i} , \acute{o} , \acute{o} , \acute{u} and \acute{u} , respectively, with i , o , \bar{o} , u and \bar{u} and hence considered a 31-letter *Hungarian* alphabet (including the space between words). The authors rewrote a random collection of articles from recent *Hungarian* newspapers conforming to the alphabet decided upon by them; then an excerpt of about 100 letters was read aloud to a group of selected subjects and they were asked to distribute stakes on the possible appearances of next letters. The first 10–15 attempts to forecast the next letter were carried out just for educating the gamblers and then the gamble began in right earnest and ended after 50 or more bets. In all the cases the committee gambling estimate due to Cover and King gave the best result. Nemetz and Simon's experiment led to the conclusion that the entropy of written *Hungarian* lies between 1.13 and 1.49 bit/letter (the average estimate yielded by this experiment is $H_\infty \approx 1.25$ bit/letter, i.e., $R \approx 75\%$). Of course, this estimate is also a preliminary one and it obviously overestimates H_∞ (i.e., underestimates R).

Recall that throughout in the foregoing we added to the number of 'letters' an empty *space between the words* (this is quite natural from the view point of telegraphy). However, it is sometimes of interest to consider also the ordinary alphabet without making allowance for space; thus, for example, we can take up the question of the information contained in one *printed* text letter. A few examples of entropy evaluations for an ordinary ('spaceless') alphabet have been presented above (cf. Manfrino [128]). It is clear that, if we drop space from our consideration, then the results deduced above undergo some modifications. Thus, for instance, it is now necessary to consider the *English* alphabet as a 26-letter alphabet, so that $H_0 = \log 26 \approx 4.70$ bits. The frequencies of individual letters also change their values and this leads to a modified value of H_1 . The new value of H_1 can be easily deduced from the early value (for an 'alphabet inclusive of space') if the average length w of the *English* words is known. Indeed, if space is considered as a zero alphabet letter, then its probability clearly equals $p_0 = 1/(1 + w)$ (on the average, one space per $w + 1$ 'letters' of a text with spaces). Moreover, the relative frequencies of all 'real' alphabet letters get

changed in the same proportion if space is included in the 'letters' under consideration (because the quantum of all 'real letters' remains here unchanged). Hence, if p'_1, p'_2, \dots, p'_n are the probabilities of 1st, 2nd, \dots , n th letter of an alphabet without a space, then corresponding probabilities p_1, p_2, \dots, p_n of the same letters with the inclusion of 'space' in the number of 'letters' are given by the equations: $p_i = (1 - p_0)p'_i, i = 1, 2, \dots, n$. Let us now write

$$-p_0 \log p_0 - p_1 \log p_1 - \dots - p_n \log p_n = H_1^{(\text{with space})},$$

and

$$-p'_1 \log p'_1 - p'_2 \log p'_2 - \dots - p'_n \log p'_n = H_1^{(\text{without space})}.$$

Then, it follows easily from the above equations that

$$\begin{aligned} H_1^{(\text{with space})} &= -p_0 \log p_0 - (1 - p_0) \log (1 - p_0) + (1 - p_0) H_1^{(\text{without space})} \\ &= h(p_0) + (1 - p_0) H_1^{(\text{without space})}, \end{aligned}$$

where $h(p_0)$ is the function defined on p. 49 (this equation was referred to in the footnote on p. 193; see also the footnote on p. 206). It is known that $w \approx 4.5$, $p_0 \approx 1/5.5 \approx 0.182$ in the case of *English* language. Using this value of p_0 and the value of $H_1^{(\text{with space})} = 4.03$ bits given on p. 193 we obtain the new value $H_1^{(\text{without space})} \approx 4.14$ bits.

The values (in bits) of the letter entropies H_0, H_1, H_2, H_3 and also the approximate estimate of the values of H_5 (or H_4) and H_8 for the 26-letter *English* alphabet, obtained by Shannon [159] with the rejection of spaces between words, are listed in the accompanying table.

H_0	H_1	H_2	H_3	H_5 or H_4	H_8
4.70	4.14	3.56	3.3	≈ 2.6	≈ 2.3

By comparing this table with that given on p. 187 we are convinced that an allowance for spaces between words in the *English* language leads to an increase in the entropy H_0 and a decrease in all succeeding entropies H_N . The fact that for all languages $H_0^{(\text{with space})} > H_0^{(\text{without space})}$ is completely obvious, since we always have $\log n > \log (n - 1)$.

Furthermore, an allowance for space increases by one number the possible outcomes of letter-guessing experiment α_1 and thus increases its degree of uncertainty H_1 , but simultaneously this allowance leads to the emergence of an additional 'letter' with an extremely large probability in comparison to others, which facilitates the forecast of the outcome of experiments α_1 and, consequently, decreases its degree of uncertainty H_1 . We see that the second circumstance

turns out to be more important in the case of the *English* language and this leads to the inequality $H_1^{(\text{with space})} < H_1^{(\text{without space})}$. Of course, the last result may not be true for all existing languages. For example, $w \approx 5.92$ for *German* (see Pratt [149]), i.e., the average word-length is here considerably larger than that in *English* and, consequently, the space probability p_0 for 27-letter alphabet is considerably smaller in *German* than in *English*. This curtails the role of the second circumstance indicated, and in fact the application of the relationships derived between $H_1^{(\text{with space})}$ and $H_1^{(\text{without space})}$ to the *German* language leads to the conclusion that here $H_1^{(\text{without space})}$ is slightly smaller than $H_1^{(\text{with space})} \approx 4.10$ bits (cf. p. 193). However, when N is sufficiently large (exceeds the average word-length), the outcome of an experiment consisting of the determination of the N th text letter with respect to the known $N - 1$ preceding letters in all those cases in which this N th letter turns out to be a 'space' is practically defined uniquely by the very structure of a language. (It is easy to understand that for large N an error in guessing the outcome of this experiment most usually takes place only when the N th letter happens to be the first, or the second letter of a new word.)† This implies that an allowance for space appreciably decreases the uncertainty of this experiment and hence if N is large, then

$$H_N^{(\text{with space})} < H_N^{(\text{without space})}$$

for every language.

It is also possible to obtain an exact relationship that connects two values of redundancy R , calculated with and without rejection of word spaces. In fact, consider two identical sufficiently long texts, which differ only in that in one of them we do not take note of spaces between words. Each text is uniquely reproduced from the other: obviously, all word spaces can be discarded in an ordinary text and it is usually quite easy to restore the spaces in a 'closely' written (without word spaces) text in a familiar language. Hence, it can be concluded that the 'total information' (the product of 'specific information', or 'information per text letter' H_∞ , by the number of letters) contained in both texts must be one and the same. But since the number of 'letters' in a text with spaces exceeds the number of 'letters' in a 'closely' written text by $(w + 1)/w$ times, where w is the average word-length (because on an average one space is required for w text letters), hence

$$H_\infty^{(\text{with space})} = H_\infty^{(\text{without space})} \cdot \frac{w + 1}{w}.$$

†This intuitively obvious statement is in good agreement with the quantitative data due to Carson [82]. According to his estimate of the numerical values of the entropies of first, second, third, . . . , letters of a word in printed *English*, the entropy of a letter decreases sharply with the increase in the number of preceding word letters,

Noting further that the probability p_0 of space equals $1/(w + 1)$ and, consequently, $w = (1/p_0) - 1$, we can rewrite this equation as†

$$H_{\infty}^{(\text{with space})} = H_{\infty}^{(\text{without space})} : \frac{1/p_0}{\frac{1}{p_0} - 1},$$

or

$$H_{\infty}^{(\text{with space})} = (1 - p_0) H_{\infty}^{(\text{without space})}.$$

However, if the total number of alphabet letters (including space) is n , then $H_0^{(\text{with space})} = \log n$, $H_0^{(\text{without space})} = \log (n - 1)$, and

$$\frac{H_{\infty}^{(\text{with space})}}{H_0^{(\text{with space})}} = \frac{H_{\infty}^{(\text{without space})}}{H_0^{(\text{without space})}} \times (1 - p_0) : \frac{\log n}{\log (n - 1)},$$

or

$$(1 - R^{(\text{with space})}) = (1 - R^{(\text{without space})}) \times (1 - p_0) \frac{\log (n - 1)}{\log n}.$$

This is the equation we need to connect the values of redundancy for a language that are obtained with and without the rejection of spaces.

Similar arguments may also be used for ascertaining the average amount of

†This result can be proved in a highly straightforward manner even without reference to the constancy of 'total information.' In fact, suppose that α_N is an experiment consisting of guessing the N th letter of a text with word spaces with respect to the $N - 1$ preceding letters. The outcome of α_N can be determined in two steps: in the first place, it is verified whether or not the N th 'letter' is a space (experiment β); if it is not a space, then we further ascertain what this letter specifically is (experiment α'_N). If p_0 is the probability of a space, then obviously we are required to carry out the second experiment α'_N only in the $(1 - p_0)$ th fraction of all cases. Hence, it follows that

$$H(\alpha_N) = H(\beta) + (1 - p_0)H(\alpha'_N),$$

where $H(\alpha_N)$, $H(\alpha'_N)$ and $H(\beta)$ are the average *conditional* entropies of corresponding experiments, given that the preceding $N - 1$ letters are known to us (see Section 2.4). If $N = 1$, then obviously $H(\beta) = -p_0 \log p_0 - (1 - p_0) \log (1 - p_0) = h(p_0)$, $H(\alpha_N) = H_1^{(\text{with space})}$, $H(\alpha'_N) = H_1^{(\text{without space})}$, and we are back to the equation derived above for $H_1^{(\text{with space})}$. However, for large N it can be considered that $H(\beta) = 0$ (the space is restored uniquely with respect to the preceding $N - 1$ letters) and $H(\alpha_N) = H_{\infty}^{(\text{with space})}$, $H(\alpha'_N) = H_{\infty}^{(\text{without space})}$; hence

$$H_{\infty}^{(\text{with space})} = (1 - p_0) H_{\infty}^{(\text{without space})}.$$

of information $H_{\infty}^{(\text{word})}$ contained in one text *word*. The zero-order entropy of one word $H_0^{(\text{word})} = \log K$ can be estimated by calculating the number of words K in any sufficiently complete dictionary of a given language; the entropy

$$H_1^{(\text{word})} = -p_1 \log p_1 - p_2 \log p_2 - \dots - p_K \log p_K$$

can be calculated with the aid of a 'frequency dictionary', indicating the frequencies (probabilities) p_1, p_2, \dots, p_K of individual words of a given language (cf. p. 186 above). However, a direct determination of the 'first-order conditional entropy' $H_2^{(\text{word})}$ demands knowledge of the frequencies of all possible two-word combinations, whose determination is practically impossible since the total number of such combinations is immensely large. The problem of calculating the succeeding 'conditional entropies' $H_3^{(\text{word})}, H_4^{(\text{word})}, \dots$, is even less tractable. Moreover, one must bear in mind that the statistical relations between individual words are frequently appreciably more rigid than those between the letters (the appearance of a word '*ENTROPY*' in a text restricts the probabilities of succeeding words more strongly than, say, the occurrence of a letter '*G*' restricts the probabilities of succeeding letters) and that these relations are considerably more 'long-range' (if the word *TOPOLOGICAL* appears at the beginning of an arbitrarily voluminous book, then this sharply decreases the probability of the occurrence of the word '*RHENOCEROS*' at its end). This creates the impression that the problem of the determination of the 'limit entropy' ('specific information') $H_{\infty}^{(\text{word})}$ must be exceedingly difficult.

Now let us associate two texts with each other, one written in the usual way by means of letters and the other 'hieroglyphic', in which a whole word is taken as a 'letter' (hieroglyphic writing is characterized by the fact that in it individual characters denote whole words). Here each of the two texts is obviously uniquely reproduced from the other, since by knowing all letters of any text we also know thereby all words occurring in it, and a knowledge of words is equivalent to knowing all the written letters. Hence here also the 'total information' contained in two texts remains the same, i.e.,

$$H_{\infty}^{(\text{word})} \times \text{number of text words} = H_{\infty}^{(\text{letter})} \times \text{number of text letters}.$$

But since the ratio of the number of letters to the number of words equals the average length of the word, we have

$$H_{\infty}^{(\text{word})} = H_{\infty}^{(\text{without space})} \times w, \text{ or } H_{\infty}^{(\text{word})} = H_{\infty}^{(\text{with space})} \times (w + 1),$$

where w is the average word-length (and hence $w + 1$ is the average number of 'letters' per word, to which is added also the space between words).

The preceding equation implies the relation

$$\frac{H_{\infty}^{(\text{word})}}{H_0^{(\text{word})}} = \frac{H_{\infty}^{(\text{letter})}}{H_0^{(\text{letter})}} \times (w + 1) : \frac{\log K}{\log n},$$

or

$$(1 - R^{(\text{word})}) = (1 - R^{(\text{letter})}) \times (w + 1) \frac{\log n}{\log K},$$

where, as above, w is the average word-length, K is the total number of words encountered in the text under consideration, n is the number of alphabet 'letters' to which is added also the space between words; here, as almost everywhere in the above, by $H^{(\text{letter})}$ and $R^{(\text{letter})}$ is understood $H^{(\text{with space})}$ and $R^{(\text{with space})}$. In particular, for the *English* language we have $n = 27$ and $w + 1 \approx 5.5$. Putting $K = 50,000$ (the approximate number of words in a moderately complete dictionary)†, we obtain

$$(1 - R^{(\text{word})}) = (1 - R^{(\text{letter})}) \times 5.5 \frac{\log 27}{\log 50,000} \approx 1.68 (1 - R^{(\text{letter})}).$$

It is thus seen that the redundancy for words is appreciably less than that for letters, i.e., 'hieroglyphic' writing is, let us say, more 'advantageous' than the customary writing by using letters. This position is closely related to the advantage from using direct long block coding of a large number of 'letters', of which we have much to say in the present chapter; words are also specific 'blocks' (such 'blocks', whose probability of occurrence is comparatively high).

It is clear that similar arguments enable us also to associate the values of the entropy (information) $H = H_{\infty}$ and the redundancy R assigned to one text letter with the same quantities determined for any other linguistic formation (*syllable*, *phrase*, *morpheme* etc.; cf. what is stated below on *phonemes*). This position explains the reasons why an overwhelming majority of information-theoretic investigations of a language start from its alphabet *letters*. In fact, a relation between the values of entropy assigned to one letter, syllable, word etc., allows us to confine the consideration to any one of these quantities; on the other hand, alphabet letters have the advantages of being familiar, uniquely defined (because for a majority of other linguistic formations like syllables, morphemes, or even words, there exist no precise definitions, excluding fully different

†Since the number of words K appears in the preceding formula under the sign of the logarithm, the inaccuracy in determining this number does not significantly influence the final result (e.g., if we put $K = 100,000$, then the factor 1.68 in the formula that follows is changed to 1.58).

interpretations of the same concept), and bounded in their number (since the 'alphabets' of words, and especially of the sentences of the language, are practically unbounded).

Let us now note that the relation between the values of $H^{(\text{letter})}$ and $H^{(\text{word})}$ may be used in *two-fold* way. This relation enables us to deduce the estimate of $H^{(\text{word})}$ from the value (supposed to be known) of $H^{(\text{letter})}$. However, on the other hand, the same relation also permits us to estimate the entropy $H^{(\text{letter})}$ by relying on the approximate values of $H^{(\text{word})}$ obtained by some method. The approximate value of $H^{(\text{word})}$ (precisely speaking, the value of first-order entropy $H_1^{(\text{word})}$) can be calculated, say, by using the so-called Zipf principle, which says that *when words of a language are arranged in the order of their frequencies* (i.e., probabilities), *the frequency of the n th probable word for all not too large values of n is found to be approximately proportional to $1/n$* . This principle was formulated and verified through analysis of a large amount of linguistic material by Zipf [179]; later it was repeatedly discussed and sharpened by several authors.† The Zipf principle has been discussed at length in Chapters 5 and 12 of [17], in Part 1 of [71] and in the papers [125], [131] where, in particular, the graphs borrowed from [179] are reproduced. These works demonstrate the applicability of Zipf's principle to texts written in different languages and having different character (say, to the text of Joyce's novel *Ulysses* and that from one of the American dailies). Shannon [159] was the first to show the usefulness of Zipf's principle in the evaluation of the first-order word entropy (and, proceeding from this, even in the approximate determination of the entropy of a letter; see pp. 186—88). Similar calculations for *Italian* language were carried out by Manfrinco [128]; for further relevant data in this direction, the reader may refer to the papers of Newman and Gerstman [137], Miller [131] and Grignetti [105].

An approximate estimate of first-order entropy $H_1^{(\text{word})}$ was obtained (with reference to the *Rumanian* language) by Voinescu, Fradis and Mihailescu (see third work of [172]) by the formula

$$H_1^{(\text{word})} = -p_1 \log p_1 - p_2 \log p_2 - \dots - p_K \log p_K.$$

Factually, however, this work is devoted to the entropy of not written but spoken language (the frequencies p_1, p_2, \dots, p_K are determined here from an analysis of magnetophone recording of answers given to a long series of standard questions by ten different subjects); hence it has been more appropriately dealt with in the next sub-section of the present chapter (see pp. 220—21). Furthermore, we note that the basic objective of the studies of Voinescu *et al.*, consisted not entirely of the determination of the value $H_1^{(\text{word})}$ for the ordinary *Rumanian*

† Thus even Zipf himself had remarked that in some cases it is more appropriate to consider that the frequency of the n th word is in fact proportional to $1/n^a$, where the constant a is close to unity, but nevertheless not exactly unity (see in this context also [6], [71], [125] and [127]).

language, but of a comparison of the value of $H_1^{(\text{word})}$ associated with the speech of healthy persons to the corresponding values associated with the speech of another ten subjects, aphasia patients (i.e., those suffering from speech disorder caused by some brain disease). Hence it also borders on a study of the statistical characteristics of different 'specialized languages', which we shall presently consider.

The data on the entropy of one letter of text, of which we spoke above, related as a rule to the 'average literary language', since *literary* texts mostly serve as the experimental input for determining entropy. Thus, Shannon [159] (working in collaboration with his wife Betti Shannon) and also Cover and King [86] analyzed the fragments from Dumas Malon's book, *Jefferson the Virginian*. Moreover, Kolmogorov and his associates used the works of Aksakov and Goncharov (see pp. 200—201). But on p. 178 it has been indicated that the occurrence frequencies of different letters may depend on the character of the considered text; exactly in the same way, the values of the entropies H_N or redundancy R will be different for texts borrowed from different sources. Moreover, any 'specialized language' (for example, a scientific or engineering text on a specific problem, business correspondence, schoolboy slang, any non-customary jargon) will, as a rule, have more than average redundancy because the number of words being used will be less and special terms and phrases will be repeated often. This circumstance is of great advantage, since it highly facilitates very fast reading of special scientific literature by the experts and even the reading of such literature in a poorly known language. Some slangs and scientific jargons may be an exception in this connection, if they are used from the especial objective of decreasing the redundancy of language. By way of example, we may mention thieves' cant, in which long and meaningful phrases may sometimes be substituted by extremely short expressions, or some recently innovated scientific jargons with enormous detailed terminology like those used in mathematics by the French school of Nicolas Bourbaki.† An even more striking example in this direction is provided by the symbolic language of modern mathematical logic, characterized by the exceptional richness of sense.

A number of authors tried to investigate the influence of the nature of the text on the values of the entropies of different orders per text letter and the text redundancy. Nevertheless till now there are available only a few results of restricted reliability in relation to this problem. For example, as already mentioned above, Manfrino's calculation [128] of letter entropy values for the *Italian* and *Portuguese* languages were carried out for three different types of text, namely that from a scientific book, a history book or novel and a newspaper. However, the entropies of orders 1, 2 and 3 only were considered by Manfrino and the values of H_N , where $N = 1, 2$ and 3, obtained for these three different types of text turned out to be very close to each other in the case of both the *Italian* and *Portuguese* languages.

More conclusive results were obtained by Newman and Waugh [138]. As already indicated, these authors calculated the approximate values of H_N for comparatively large N with the aid

† A more popular example is analyzed in [135], which has been already mentioned before (p. 191).

of the method due to Newman and Gerstman [137], which is unfortunately not much dependable. However, this made it possible for them to obtain crude estimates of all the entropies H_N up to the twelfth order (and the corresponding redundancy of the twelfth order $R_{12} = 1 - H_{12}/H_0$). These estimates of Newman and Waugh were derived by evaluating three 10,000-letter excerpts sampled from English *Bible*, the writings of American philosopher and psychologist William James and the modern magazine, *Atlantic Monthly*. According to the results obtained, the values of the entropies of a few lowest orders do not differ much for these three different types of text, but those of higher order entropies and redundancy differ considerably. The prose of the *Bible*, the simplest of the three samples, is characterized by the lowest value of the entropy per letter and the highest redundancy (and also by the lowest value of the average word length w ; for the data on the value of w , see pages 187 and 196), while the highest value of H_N and w and the lowest value of R are attained for the most terse prose from the *Atlantic Monthly*; the writings of William James range between the two other texts in all these respects, though they are much closer to the *Atlantic Monthly* than the *Bible*.

Newman and Waugh's method was slightly modified by Carterette and Jones [83] to study the entropies of a few lowest orders and estimate the redundancy in children's graded reading books. Since the text difficulty must increase within a series of children graded readers, it is natural to assume that the entropy values would increase and redundancy would decrease with increase in the reader's level. To verify this assumption, Carterette and Jones analyzed children's reading texts at levels 1, 2, 3 and 5 and compared them with each other and with three types of adult texts which were investigated by Newman and Waugh. A 28-letter alphabet was chosen in Carterette and Jones' study by adding to the 'customary' 26 *English* letters the period (the end of a sentence) and the space (the end of a word), but this circumstance is of minor importance. The results obtained by them [83] are in good agreement with the expectations: they show that the text redundancy decreases progressively from the *First reader* to the *Atlantic Monthly*, with the *Bible* being close to the *Third Reader* in this respect and the *Fifth Reader* approaching William James' writings intended for adults.

A mention has already been made of letter-guessing experiments carried out by Weltner [173] and Piotrovskii and his coworkers (see [147], [143], [146] and [148]) for various types of texts. In particular, Weltner's book contains the figure and the detailed table showing the estimates of the entropy (per text letter) for a great variety of texts including poems, a series of prose texts (short stories, novels by various authors), excerpts from two different newspapers, scientific texts in different fields and a number of usual and programmed textbooks. All estimates were deduced from the results of letter-guessing by the same group of subjects (students of a teacher's college), which demonstrated considerable differences between the entropies of different texts. However, Weltner did not attempt to clarify whether these differences were statistically significant in all the cases or stemmed from the experimental errors.

Piotrovskii and his coworkers studied three different types of texts: conversational language, literary texts (i.e., fiction) and various business texts (including engineering and scientific writings). The results (see [143], [146] and [147]) related to the information-theoretic characteristics of three types of texts in the *Russian* and *French* languages and the averaged results (for the language at large) are listed in the table† on p. 212. In complete agreement with what has been stated above they show that the redundancy of 'business texts' is appreciably greater than both the 'average redundancy' and the redundancy of literary texts. However, the redundancy of conversational language is found to be slightly lower than the averaged redundancy—in principle, this may be due to the 'liberty' permissible in conversational language which often leads to violations of strict constraints dictated by subtleties of 'style' and rules

†In the references cited, there are some discrepancies between the values of redundancy R and the entropy H . However, in the table on p. 212 the values of R have been brought in conformity with the values of H taken from the same sources.

TABLE

	$H = H_{\infty}$ (in bit/letter)		R (in per cent)	
	<i>Russian</i> <i>Language</i>	<i>French</i> <i>Language</i>	<i>Russian</i> <i>Language</i>	<i>French</i> <i>Language</i>
Language at large	1.37	1.40	72.6	70.6
Conversational language	1.40	1.50	72.0	68.4
Literary texts	1.19	1.38	76.2	71.0
Various business texts	0.83	1.22	83.4	74.4

of grammar. In [148], similar results are presented for seven languages (*Russian, French, English, German, Polish, Rumanian and Kazakhian*) and an attempt is also made to invoke some rather crude procedures of mathematical statistics to test whether the redundancy differences between various texts are real (statistically significant), or fictitious (generated by experimental errors). These tests enabled the authors to conclude that the considerable difference obtained between the redundancy of literary texts, or conversational language and that of business texts is statistically significant for all the studied languages, but a small difference obtained between the redundancy of literary texts and that of conversational language is probably due to the experimental errors.

The study by Smirnov and Yekimov [163] bears a more special character. These authors investigated a sample from *Russian* telegraphic texts, the size of which was about 15,000 letters, using Shannon's letter-guessing method (and its variant due to Kolmogorov; see p. 198 et seq.). The main result obtained by Smirnov and Yekimov says thus : $H^{(\text{telegr. Russian})} \approx 1.4H^{(\text{literary Russian})}$. This result is obviously connected to the deliberate decrease in the redundancy of telegraphic text (say, owing to the omission of conjunctions and other 'evident' words).

The other highly specialized language, namely, the so-called 'control tower language' of radio communications between the air traffic controller at airport and the aircraft pilot in air, has been studied by Frick and Sumbly [97] as well as by Fritz and Grier [98]. The radio communications considered in these works are naturally quite standard in their form and confined to a few limited, constantly recurring topics. Hence, it is no wonder that the redundancy of the corresponding language (estimated either by means of 'guessing experiments' or through a direct statistical study of the collection of a few standard sentences of which these communications are made up) is found to exceed considerably the redundancy of average 'literary language.' In the particular, by confining themselves further to a very restricted class of messages transmitted by an airport 'control tower operator' to the pilot landing a plane, Frick and Sumbly obtained for the redundancy a value close to 96% (almost the same redundancy value close to 93% can be deduced from the results obtained by Fritz and Grier). The abnormally large redundancy has here a completely transparent justification—because of the difficulty in receiving the message (due to the aircraft noise), a reduction in redundancy may lead to erroneous reception, foreboding, in the considered case, disastrous (even tragic) consequences. Hence the high redundancy is here necessary for air traffic security.

The fact that a 'specialized language' is characterized by high redundancy is used when, for instance, one constructs specific codes for the business correspondence of a large firm. At present, such codes are developed with the indispensable participation of information theory specialists, and the presence of many oft repeated standard words and sentences in the firm's correspondence facilitates the increase of the code efficiency considerably.

Let us now examine the interesting but so far little studied question of the differences in the language redundancy of different *literary* texts. It can be presumed that different literary genres are distinguished by different redundancies, related to the style intrinsic specifically to this type of composition; it can also be conceived that even within one literary composition in different fragments (dialogue, description etc.) the redundancies are different. High redundancy may characterize the hackneyed, stereotyped language of a literary composition, but can also serve as only an evidence of the leisurely style of an author (thus, a high redundancy is detected in the experiments mentioned on pp. 200—201 on the letter entropy estimation in Goncharov's *Literary Party*, written in a placid and flowing language spelling out a large number of quite obvious details). Low redundancy may bear testimony to the richness and brilliance (unexpectedness, unconventionality) of a literary language (possibly, Faulkner's language can be cited as an example here); an extremely low redundancy in the language of a literary composition is invariably interpreted as a deliberate complication of the language, however (cf. *Finnegans Wake* by Joyce). Still lower redundancy will have the sort of 'obscurity' that was used by the Russian poet Khlebnikov at the turn of the century and became popular among a number of Western poets after the second world war (recall that zero redundancy characterizes the 'sentence' mentioned on p. 178, which can hardly be considered as a distinctively 'nice' literary form).

The allied problem here is that of comparing the redundancies of prose and poetic language, widely discussed in the sixties (see [116], [129], [165] and a number of papers in the collection [117]; see also the articles of Dolezel and of Nicolau, Sala and Roceric included in [128]). It is clear that the poetic form characterized by a specific rhythm and rhymes imposes on a language certain additional restrictions, i.e., raises its redundancy. An attempt can also be made to estimate quantitatively, say, the impact of the rhythm of a verse, by determining the quantity of word combinations satisfying a given rhythmic plan, and comparing it with the entire store of meaningful word combinations (in the determination of such store, it is convenient to use as a base a dictionary of prose compositions by the same author).† The impact of rhymes is slightly more intricate to calculate, but rough estimates are completely possible here, too. The approximate estimates deduced by Kolmogorov for the classical *Russian* tetrametric iambic verse (for example, Pushkin's *Evgenie Onegin* written in this verse)†† show that the fulfilment of the requirements imposed on the poetic form

†See, for example, Kondratov [116] in which low order entropy is calculated, determined by the *Russian* poetry of a definite rhythmic plan and by the *Russian* prose (scientific, business-like, fiction, colloquial) texts (in bit/syllable); cf. also Lüdtke "A comparison of metric plans with respect to their redundancies" in [117].

††The tetrametric iambic verse is characterized by a stanza, which theoretically consists of eight uniformly alternating accented and unaccented syllables (in practice some accents are sometimes shed).

reduces the 'uncertainty' H_∞ for one text letter by a quite appreciable amount, whose order compares with the half value of H_∞ calculated for an 'averaged literary' text. In fact, the corresponding letter-guessing experiments carried out by Kolmogorov also show that, for a 'poor' verse (in which the decrease of information contained in a letter is not compensated for by emotional stimulation, brilliance of speech and richness of language characteristic of 'good' poetry), the 'limiting information' H_∞ per text letter is essentially less than (approximately half of) the value of H_∞ determined for classical *Russian* prose.† However, in the compositions of many eminent poets, the decrease in the information content of one text letter, related to the fulfilment of known formal rules, is apparently compensated for to a great extent by the enhanced radiance and unconventionality of language. Therefore, it can be well expected that here the redundancy of the language has the same order as that of a prose literary text.

The impact of various factors related to literary style on the value of the entropy and redundancy of language is considered by Paisely [140]. He made use of the method due to Newman and Gerstman [137] and Newman and Waugh [138], which is not quite reliable but it enabled him to analyze 39 different *English* excerpts and compare the entropies among them. The compared excerpts include: (a) two poetic translations from Homer's *Iliad* due to different authors; (b) four translations of two different passages from the same *Iliad*, and also four (modern) translations of two passages from a chapter of Matthew's *Gospel* (in both cases the selected passages differ considerably in content); (c) four prose and four poetic translations from *Iliad*, and (d) nine different translations from Matthew's *Gospel* relating to different periods. In a number of cases analyzed by Paisley the differences among the entropy values turned out to be noticeable, and some general regularities could also be noted here (such as the progressive decrease of redundancy in literary texts with the time of their writing approaching more nearly modern times). However, all these inferences are still not quite dependable and call for further verification.

The studies [160] devoted to a number of *Indian* languages are of a nature similar to those mentioned above. In these papers, the values of the entropy calculated for texts of different character (say, prose and poetic) and different times of writing are also enumerated. Some of the results obtained in [160] definitely have something in common with the results obtained by Paisley on material written in the *English* language. However, a comparison is rendered difficult here due to the substantial difference between the *English* and *Indian* alphabets (see the discussion on pp. 197—198).

Of the works more directly related to the comparison of statistical characteristics between prose and poetic language (the question is not lost sight of in [140] and [160]), the foremost to mention are the investigations of Doležel and of Nicolau, Sala and Roceric (see [128]) on an evaluation of the entropies of various orders for *Czech* and *Rumanian* prose and poetic language, and even for individual prose writers and poets. However, the preliminary estimates obtained by these authors clearly need further sharpening. Marcus [129] made a bold attempt to carry over to poetry the relationship between the physical concepts of 'entropy' and 'energy': on this basis he considered some results contained in the studies of Nicolau, Sala and Roceric, concerning the calculation of the entropy for M. Eminescu's compositions, relating to various periods of the poet's creative works. Tarnóczy's paper [165] has a more special character; it contains the evaluation of a number of information-theoretic characteristics of *Hungarian* prose and poetry.

†The short novel *Duel* by the Russian writer A. I. Kuprin was compared to a poem of quite poor quality printed on the reverse of one of the sheets of a torn-off calendar.

Finally, let us note that well-justified doubts were and are still entertained about the application itself to literary texts (unique by the very definition!) of conventional information-theoretic ideas arising in connection with purely applied problems of communication engineering. In fact, information theory ignores the question of the subject matter of the transmitted message and relies only on purely statistical concepts (e.g., on concept of letter frequencies in a 'statistical ensemble' of all 'average texts' of a given language; but what can be made out of the notion of a 'statistical ensemble' consisting of Shakespeare's tragedies, or Pushkin's poems?). These considerations led Kolmogorov (see [15]) to undertake an extensive formulation of the problem of the possibility of various approaches to the very notion of 'amount of information' and suggest a 'pure combinatorial' approach to this concept, in particular in application to a study of the language entropy and, especially, the entropy of literary texts.

The essence of *combinatorial* approach to the determination of entropy consists of the following. The *Shannon entropy* H per text letter can be determined subject to the condition that, for an n -letter alphabet, the number of different N -letter texts (where N is sufficiently large) satisfying the given statistical restrictions is not equal to the number $n^N = 2^{\log^n \cdot N}$ ($= 2^{H_0 N}$), to which it should be equal if we have the right to choose *any* collection of N sequential letters, but is equal only to $M = 2^{HN}$ (see pp. 55—56 and 168—169). In accordance with this, by using the notion of 'intelligible' text, we can determine the entropy H as

$$H_{\text{comb}} = \lim_{N \rightarrow \infty} \left(\frac{1}{N} \log M(N) \right),$$

where $M(N)$ is the number of all possible *intelligible* texts of length N . This definition no longer depends on any probability-theoretic concepts.

In striving to estimate numerically the value of the 'combinatorial entropy' H_{comb} , the number $M(N)$ may be estimated by means of the calculation of the number of possible *extensions* of the text. Expressly, suppose that $*$ is a 'blank' word that contains no letter at all; further, denote by $l(*a_1a_2 \dots a_k)$ (or by $l(a_1a_2 \dots a_k)$, where a_1, a_2, \dots, a_k are some letters of the language under consideration), the number of all possible 'intelligible one-letter extensions' of a sequence of letters $a_1a_2 \dots a_k$, i.e., the number of such letters x that the fragment $a_1a_2 \dots a_kx$ can be extended up to an intelligible text. In this case, the value

$$\hat{M}(n) = l(*)l(*a_1a_2)l(*a_1a_2) \dots l(*a_1a_2 \dots a_{N-1}),$$

averaged over the number of letter sequences, can be considered as an estimate of the quantity $M(N)$ we are interested in.

What has been stated above paves the way for a purely combinatorial evaluation of the entropy and redundancy of a 'grammatically correct' text. The earliest efforts in this direction are traced to Kolmogorov and his coworkers (see the first paper of [15]), in which the number of possible text extensions is determined with the aid of the list of words entered in S. I. Ozhegov's *Dictionary of the Russian Language*. The estimate $H = (1.9 \pm 0.1)$ bit/letter obtained here naturally appreciably exceeds the bound on the entropy of 'literary texts' indicated on p. 201 (since the 'degree of uncertainty' of a literary text letter is by no means bounded only by the requirements of grammatical correctness). Unfortunately, a more detailed exposition of these investigations as well as the results of similar studies, commenced in Leningrad by R. A. Zaidman, has not yet been published.

4.3.2. Spoken Language

We now pass on to the problem of the entropy and information contained in *spoken* language already touched upon on pp. 211—212. It is natural to

think that all statistical characteristics of such speech depend considerably more on the choice of the speakers and on the character of their talk than that observed in the case of written language; in fact, written language is, as a rule, more 'uniform' than spoken language. Though according to the data due to Piotrovskii and his colleagues, 'on the average' the entropy of spoken language is slightly higher than the entropy of written texts, this is undoubtedly not so for certain types of speech (cf., say, the example of 'control tower language' in p. 212). The lower value of the entropy of speech can be explained by the fact that in conversation a few words are often repeated many times (there being least concern about the 'elegance of style') and also many 'superfluous' words (i.e. having no information content) are frequently added; this happens both for facilitating the understanding of speech and then just allowing the speaker some time to think about what he desires to say next. In particular, the redundancy of speech is very large when the level of noise is high (say, in the humming of an airplane, in a compartment of an electric train or subway), and also for conversations between drunkards, persistently repeating one and the same words and expressions (as a rule, not very sophisticated ones); the latter is explained by the fact that in this case not only the understanding but also the pronunciation of speech is difficult.

By determining the average number of letters pronounced per unit of time, it is possible to estimate approximately the amount of information conveyed during a conversation in 1 sec; usually it is of the order of 10 bits (this information amount naturally depends strongly on the 'conversation speed' which can be varied quite significantly: 'very rapid' speech is almost five times faster than 'very slow' speech).[†] This is in agreement with physiological acoustic data, which enables us to estimate the total number of 'distinguishable sounds' pronounced by a person in unit time (see Miller [131]).

However, this estimate of the information transmission rate for a conversation takes account of only the 'semantic' information, which is related to the meaning of the speech and can be extricated also from a write up of the stated words. In fact, a real speech always contains, in addition to this, further sufficiently significant supplementary information, which the speaker communicates sometimes voluntarily but sometimes also directly contrary to his own desire; this supplementary information may even contradict the 'semantic information' but in such cases it deserves, as a rule, a greater confidence. Thus, from a conversation we can judge the temper of a speaker and his attitude to what has been stated; we can recognize the speaker, even if it is not indicated to us by any other source of information (including here also the 'meaning of

[†]We are speaking here obviously not of conversation with exceptionally high redundancy, of the sort discussed above; thus, in the case of parleys between the pilot and the air controller at an airport, the information transmission rate does not exceed 0.2 bit/sec, i.e., it is much smaller than that for extremely slow conversation on general topics.

speech'); in many cases we can determine the birth place of a person unknown to us by his pronunciation (the latter factor plays an important role in the opening act of Bernard Shaw's play *Pygmalion*); we can evaluate the loudness of speech, which in the case of voice transmission through a communication channel (telephone, radio) is determined in the main purely by technical characteristics of the transmission channel, etc. A quantitative evaluation of all this information is a highly complex problem, which demands a considerably deeper knowledge of language than that available at the present time; in particular, this requires vast statistical data of a great variety, which is almost completely lacking so far.

An exception in this respect is the comparatively restricted problem of the so-called 'insistence stresses' emphasizing individual words in a sentence. These stresses also carry a definite information load, which (for the particular case of telephone conversations in *English*) can be estimated quantitatively. The statistical data required for this were obtained by Berry [76], who analyzed a number of 'typical *English* telephone conversations'. His data show, in the particular, that the stress is usually put on the most rarely used words (that, however, is quite natural, since it is clear that anyone will hardly put stress on the most common words, say, prepositions, articles, or conjunctions). If we denote by q_r the probability of finding a definite word W_r stressed, then the average information contained in knowing whether that word is or is not stressed is given by

$$-q_r \log q_r - (1 - q_r) \log (1 - q_r).$$

Suppose now that p_1, p_2, \dots, p_K are the probabilities (frequencies) of all words W_1, W_2, \dots, W_K (here K is the total number of all words used; the probabilities p_1, p_2, \dots, p_K , playing a basic role throughout the language statistics, may be found in the so-called 'frequency dictionaries', see pp. 186 and 207). In such a case, for the average information H contained in the insistence stress, we can set up the formula

$$\begin{aligned} H = & p_1 [-q_1 \log q_1 - (1 - q_1) \log (1 - q_1)] \\ & + p_2 [-q_2 \log q_2 - (1 - q_2) \log (1 - q_2)] \\ & + \dots + p_K [-q_K \log q_K - (1 - q_K) \log (1 - q_K)]. \end{aligned}$$

By substituting here Berry's data, Mandelbrot [126] calculated that the average information, which we obtain by ascertaining on which words the insistence stresses are put, is approximately of the order of 0.65 bit/word in the case of the *English* language.†

†This calculation was set forth in Mandelbrot's paper presented at the Third Symposium on Information Theory held at London in 1955. The paper was withdrawn from the Symposium proceedings, but included in the Russian translation of these proceedings [126]. Mandelbrot's related paper in English (see [126], second paper) contains [on p. 77] the simplest form of Berry's law yielding the stated calculation but does not spell out the calculation itself.

As to the generally diverse 'unsemantic' information contained in speech, the existing data allow us to give only a quite rough and incomplete estimate of its total quantity. Such an estimate was obtained by the German scientist, Küpfmüller in his interesting study [118] of spoken and written *German* language, which has already been referred to in the foregoing. Küpfmüller did not even make an attempt to take account of the intricate statistical regularities of intonation, tone of voice, and other peculiarities of speech. His work is essentially restricted only to estimation of the 'zero-order entropy' H_0 related to the number of different possibilities, and is then offered as a rough guide on the assumption that the corresponding redundancy is equal to 50%. Together with the information given by intonation, Küpfmüller has estimated separately the information connected to the individual characteristics of the voice of a speaker and has also evaluated the information conveyed by the loudness of the speech; the sum of the three quantities obtained here has been associated with the 'semantic' information contained in the same speech. For an evaluation of the total number of identifiable degree of loudness and the total numbers of 'speech melodies' (the types of intonations determined by a small variation of the basic frequency of sound oscillations), the physiological acoustic† data have been given; the total number of individual voices discernible by a person is roughly determined, so to say, 'by eye'. It is natural that Küpfmüller's estimates of the 'total number of possible outcomes' obtained in this way cannot make any claim to a high precision; however, since the information is determined by the logarithm of this number, even a rough estimate enables us to calculate the amount of information to a quite reasonable accuracy (clearly, when the total number of outcomes is of the order of 1000, then, for the information to be twice overestimated, it is necessary that this number of possibilities be increased 1000 times!). These calculations led Küpfmüller to conclude that the supplementary information contained in the intonation, loudness, and peculiarities of individual voices in normal conversation must not be greater than 75% of the 'semantic' information; in quite rapid and extremely slow speech it forms, respectively, not more than 30% and 150% of the 'semantic' information. (The substantial difference between the three values may be explained partially by the fact that in rapid speech different voices are considerably less discernible and different intonations are much less distinguishable.)††

†Seemingly, the loudness and intonation may be varied in a continuous manner, so that infinitely many different possibilities must be available here. In reality, however, the human ear distinguishes only a finite number of different degrees of loudness and a finite number of intonations; we shall have more to say about this in detail below (see Sec. 4.3.4).

††Apparently, this position is due to the fact that the nerve channels leading from the hearing organs to the brain may transmit during a unit of time only a fixed amount of information (see pp. 249-251). Hence an increase in the 'semantic' information transmission rate invariably implies a decrease in the transmission rate of other types of information over the same channel.

In Küpfmüller's work the values of the 'specific' entropy and information of speech related to one pronounced letter are also given. Factually, however, these values have only a conditional character (they are needed just for a comparison of speech with written language). In fact, during a conversation individual letters are never uttered, but only sounds are pronounced, which differ substantially from letters. Hence it is necessary to regard an individual sound, a *phoneme*, as the basic element of speech (in the same sense that a letter is the basic element of written language). Meaningful speech is made up of phonemes in exactly the same way that meaningful written language is composed of letters. Hence, in the transmission of speech over a communication channel we have only to observe that all phonemes are transmitted correctly. If it is achieved, then the meaning of the entire speech will also be conveyed correctly, i.e., no part of the 'semantic' information will be missed. This leads to the result that in all cases, when we are interested only in the transmission of the 'semantic' information of speech (a majority of cases are so), our concern is primarily focused not on the entropy and information of a 'pronounced letter' (which is a purely conventional notion), but on the entropy and information of one actually pronounced phoneme.

The list of phonemes for a given language is obviously not identical with the list of alphabet letters. The total number of phonemes considerably exceeds the number of letters, since one and the same letter can be sounded differently in different cases (for example, the pronunciation of a vowel depends substantially on whether or not it is accented; one and the same consonant can be pronounced with hard and soft sounds and so on). It is necessary to bear in mind here that even if in relation to the number of alphabet letters different view points may be possible (cf., for example, the footnote†† on p. 192 related to *European* alphabets using *Latin* letters and the discussions on pp. 194 and 203 on the *Russian* 'telegraph alphabet' and *Hungarian* alphabet, respectively), then with respect to a 'phoneme alphabet', concerning the very definition of which (see, for example, Cherry [6] or Uspenski [170]) there is so far no consensus among linguists, the differences between various authors are inescapable. Some preliminary results about the phoneme statistics and phoneme entropies of the *English* spoken language have been obtained by Black and Denes (see [77]). The former calculated the entropies H_0 , H_1 and H_2 for one phoneme by statistical data related to a collection of one- and two-syllable *English* words (which obviously still does not characterize the entire *English* language), the number of phonemes considered being 41. The latter author determined the relative frequencies of phonemes and all their pair combinations (phonemic 'digrams') by the data related to an 'average *English* language', and by taking the number of phonemes as 45 (the entropy H_1 of one phoneme digram, as it follows from the data due to Denes, is given in [93]). Similar statistical results on the phonemes and phonemic digrams of *French* language were published by Haton and Lamotte [106]. The German scientist Endres [93] made an effort to

evaluate approximately the total redundancy of one phoneme of *German* and *English* speech by using a spectrogram of phonemes (giving the representation of a phoneme in the form of a figure on a plane) and then applying rough methods to determine the redundancy of the plane figures allied to those used in the concluding portion of [112] (which shall be further elaborated in pp. 239–241) for an estimation of the redundancy of the letter figures in a typescript text. According to his data for both languages, the redundancy of phonemes is close to 80–85% (that is, it does not differ much from the redundancy of the letters of a written language). The American scientists Cherry, Halle and Jacobson [84], who also made use of the findings of a number of Russian linguists, selected 42 different phonemes in the *Russian* language. They calculated the frequencies of individual phonemes (and also various phonemic ‘digrams’ and ‘trigrams’) by using mainly quite obsolete and incomplete data given by the well-known Russian philologist, A. M. Peshkovskii [142].† Starting from these data, they determined the values of the ‘maximum possible entropy’ $H_0 = \log 42$ for one phoneme, the first-order entropy $H_1 = -p_1 \log p_1 - p_2 \log p_2 - \dots - p_{42} \log p_{42}$ (where p_1, p_2, \dots, p_{42} are relative frequencies of different phonemes), and the second- and third-order ‘conditional entropies’ H_2 and H_3 (defined in exactly the same way as for the written language). The results obtained (in bits) [84] are listed in the accompanying table.

H_0	H_1	H_2	H_3
$\log 42 \approx 5.38$	4.77	3.62	0.70

It is instructive to compare these values with the values of the letter entropies H_0 , H_1 , H_2 and H_3 given on p. 194 for the *Russian* written language (see also p. 196). The comparison shows that if only the data in [84] are justified,†† then the decrease of a number of the conditional entropies for phonemes takes place appreciably more rapidly than in the case of the written text letters.

A study of low-order entropies in the *Rumanian* speech (and a comparison of the data obtained with that related to the written language) has been carried out by Fradis, Mihăilescu and Voinescu [96]. Let us finally mention the papers of Voinescu, Fradis and Mihăilescu [172], which are devoted to a comparison of the information-theoretic characteristics (the entropies H_1 and H_2 for one

†A considerably more extensive study of the frequencies of individual phonemes and their pair combinations (based on vastly extensive modern material) has been carried out in the Department of Phonetics of Leningrad University (see Zinder [178]). In this investigation the total number of phonemes is taken as 48 (in the first place, at the expense of a more detailed demarcation of vowel sounds).

††Unfortunately, [84] does not give an indication of the exact volume of the material used for the frequency determination of different phonemes and their binary and ternary combinations. Hence, it can be apprehended that the value determined of H_3 is strongly understated because of the insufficiency of statistical data (cf. this footnote on p. 228).

phoneme, the difference $H_0 - H_1$, and also the entropy $H_1^{(\text{word})}$, see p. 209) of the speech of healthy persons and that of the aphasic persons (i.e., those suffering from some brain disorder affecting the speech). It turned out here that the entropies H_1 , H_2 and $H_1^{(\text{word})}$ all assume appreciably lower values for the speech of an aphasic patient than for that of a healthy person (i.e., the redundancy of speech is increased here considerably) and, in addition, the stated entropies, as a rule, also differ more sharply for different aphasic persons than for different healthy persons (the especially sharp character of the indicated phenomenon was observed in application to the quantity $H_1^{(\text{word})}$ which essentially depends on the size of the speaker's vocabulary and the extent of uniformity with which words from this vocabulary are used by him).

With the aid of the arguments employed in the foregoing for the determination of the redundancy $R^{(\text{word})}$, the relation between the redundancies of spoken and written languages can also be established. The fact that speech can be written down and written languages can be spoken implies that the 'total information' contained in a specified text† does not depend on the form, whether spoken or written, in which this text is presented. Hence

$$H_{\infty}^{(\text{letter})} \times \text{number of letters} = H_{\infty}^{(\text{phoneme})} \times \text{number of phonemes}$$

(see p. 207). Consequently, it follows that

$$H_{\infty}^{(\text{phoneme})} = H_{\infty}^{(\text{letter})} \times \omega,$$

where ω is the average number of letters per phoneme ('the average phoneme length'). The quantity ω is an important statistical characteristic of a language, which connects the spoken and written languages. From the preceding formula it also follows that (cf. pp. 206 and 208)

$$\frac{H_{\infty}^{(\text{phoneme})}}{H_0^{(\text{phoneme})}} = \frac{H_{\infty}^{(\text{letter})}}{H_0^{(\text{letter})}} \times \omega : \frac{\log k}{\log n},$$

or

$$(1 - R^{(\text{phoneme})}) = (1 - R^{(\text{letter})}) \times \omega \frac{\log n}{\log k},$$

where k is the total number of phonemes, and n is the number of letters; here it is natural to take $R^{(\text{without space})}$ for $R^{(\text{letter})}$. However, the difficulty encountered in the use of this equation is the absence of statistical data, which could permit the determination of the quantity ω (even with regard to the number

† Apparently, in the case of speech only the 'semantic' information contained in it is considered (see p. 216).

of phonemes, there is so far no consensus of opinion among philologists).†

4.3.3. Music

A study of the same sort can also be carried out with respect to *musical messages*. It is natural to think that there is a quite strong bond among the sequential sounds (i.e., sequential note symbols) of a given melody. Some note sequences that are more melodious than others occur more frequently in musical compositions than the other ones. If we write out randomly a number of notes, then the information contained in each note of this entry will be the largest; however, from the viewpoint of music such a chaotic sequence of notes will be of no value. In order to obtain a tune pleasing to the ear, it is obviously necessary to insert in our sequence a definite redundancy. It may, however, be apprehended in this connection that, in case the redundancy is too large so that the succeeding notes are defined almost uniquely by the preceding ones, we obtain a most monotonous and uninteresting piece of music. Then, what is the redundancy under which 'pleasing' music can be obtained?

It is highly likely that the redundancy of simple tunes be of the same order as the redundancy of intelligible speech. It would be of great interest to study quantitatively the redundancy of various forms of musical compositions or compositions by various composers. Unfortunately, at present we have very little concrete data of this sort. One of the earliest results in this direction was obtained in 1956 by Pinkerton [145], who analyzed from the standpoint of information theory an album of popular American nursery rhymes. For simplicity it was assumed in this work that all sounds are within the range of one octave; furthermore, since the so-called chromatic scales do not occur in the considered tunes, all these tunes may be reduced to seven basic sounds: *do, re, mi, fa, sol, la* and *si* (which correspond to the white keys on a piano). All the analyzed songs were set up as a sequence of the 'basic elements', each with a range of one beat (an eighth note). To the seven notes of an octave additional eighth 'basic element' *O* was added for signifying rest or holding of a note for more than one beat. Thus, the 'maximum possible entropy' H_0 of one note is here given by

$$H_0 = \log 8 = 3 \text{ bits.}$$

By calculating the frequencies (probabilities) of individual notes in all 39

*By associating *English* phonemes with the 43 phonetic symbols used in Anglo-Russian dictionaries, widely prevalent in the USSR, we can determine approximately the 'average phoneme-length' ω by a comparison of the length of *English* words written in letters and their phonetic transcriptions. Then, we obtain $\omega \approx 1.2$, yielding

$$(1 - R^{(\text{phoneme})}) = (1 - R^{(\text{letter})}) \times 1.2 \frac{\log 26}{\log 43} \approx 1.04 (1 - R^{(\text{letter})}).$$

analyzed tunes, Pinkerton found that

$$\begin{aligned} H_1 = & -p(O) \log p(O) - p(do) \log p(do) - p(re) \log p(re) \\ & - p(mi) \log p(mi) - p(fa) \log p(fa) \\ & - p(sol) \log p(sol) - p(la) \log p(la) \\ & - p(si) \log p(si) \approx 2.73 \text{ bits;} \end{aligned}$$

here, for example, $p(do)$ denotes the probability of the note *do*. By applying the probabilities for combinations of two notes determined by Pinkerton, the conditional entropy H_2 can also be calculated; it turns out to be close to 2.42 bits. (However, let us note that Pinkerton's paper contains only the somewhat averaged values of two-note combination probabilities, so that the obtained value of H_2 is overstated.) It is clear that by means of the values of H_1 and H_2 alone there is very little that can be stated about the degree of redundancy of the considered melodies (it can only be said that obviously it is *appreciably higher* than $1 - (2.42/3) \approx 0.2$). Some indirect data that verify this conclusion are given below.

Even before the appearance of Pinkerton's paper, the work of F. and C. Attneave, calculating the frequencies of individual notes and two-note combinations in a number of American cowboy songs, was reported at the Conference on Information Theory held in London in 1955. A considerably more detailed study of this sort was accomplished in 1957 at the Computer Department of Harvard University (see Brooks *et al.* [80]). Here excerpts from 37 hymn tunes of different composers and periods of origin, having the same metric structure, were analyzed. The use of a high speed electronic computer enabled the authors to dispense with the simplification that consists of referring all notes to one and the same octave; the distinct 'basic elements' considered here were all the notes of the four octaves in the chromatic scale (including also the five intervening sounds, corresponding to the black keys of the piano). Thus, Brooks *et al.*, considered in all 49 distinct elements, and they also included here the special notations for sounds extended from the preceding time interval. The unit of duration of one basic element was again chosen as an eighth note, since shorter notes were not encountered in any of the considered hymns.

Brooks *et al.*, calculated with the aid of modern computers the frequencies of all individual 'basic elements' and all combinations of two, three, . . . , eight such adjoining elements. The results obtained yield in principle the possibility of setting up approximate expressions for all conditional entropies from H_0 , H_1 , H_2 up to H_8 , inclusive. Truly, it is necessary to bear in mind in this connection that the statistical material used (consisting of 37 small excerpts from different hymns) is surely inadequate for obtaining any reliable estimate of the probabilities of combinations of a large number of notes; hence, the values of higher order entropies (the entropies H_5 , H_7 and H_8 in every case) determined in this way have little validity. Nevertheless, the values of the first few condi-

tional entropies may be of positive interest; hence we only regret that the authors did not produce in [80] the results of the corresponding calculations (and they also adduced no such data as would permit an estimate of the corresponding entropies).

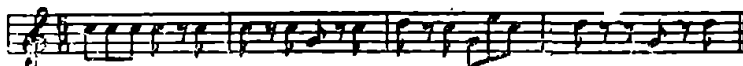
A similar analysis of the melodies of the noted American composer Stephen Foster (1826-1864) was carried out (although on a modest scale) by Olson and Belar [139]. These authors considered the 11 most popular songs of Foster, and by putting the musical scale on the basis of 12 different notes (covering one and a half octaves), they calculated the frequencies (i.e., the empirical values of the probabilities) of each individual note and all possible groups of two and three succeeding notes. It is clear that starting from the data obtained it is possible to estimate without difficulty the entropies H_0 and H_1 and even the conditional entropies H_2 and H_3 for one note in Foster's songs (although this was not accomplished in [139] either). Further information on the studies of musical composition statistics may be found in Zaripov's book [177], which also contains an extensive bibliography.

Examples of direct valuations of the information-theoretic characteristics of different musical compositions are available in the papers of Youngblood [176], Cohen [85], Siromoney and Rajagoplan [162], Hiller and Beauchamp [107], Roland [153] and some others (see also a review of this topic given in Chapter 13 of [17]). Thus, for example, in [85] (in which the results due to Youngblood and Brawly are also used) the values of the entropies H_1 and H_2 and the corresponding redundancies $R_1 = 1 - (H_1/\log n)$ and $R_2 = 1 - (H_2/\log n)$ of the first two orders, related to one note, are calculated and compared among themselves for the nineteenth century musical material of individual romantic composers (Schubert, Mendelson, Schumann) and the *German* romantic music at large, and also for a collection of Catholic religious hymns and modern American rock and roll music. In [153] the values of redundancies for the classical music of Haydn and modern music of Schönberg are compared (it is natural that in Schönberg's music the redundancy is found to be less than that in Haydn's music). In [107] some results are given of the analysis of one of the compositions due to Webern, which is similar to Schönberg's compositions, and in [162] the values of H_1 are calculated for a number of compositions of South Indian (*Karnatic*) music of the eighteenth and nineteenth centuries. In [85] and [107] are adduced also some data with respect to the 'rhythmic redundancy' of different musical compositions (similar to the redundancy of 'metric rhythm' in verse). However, so far all the estimates obtained concerning the information characteristics of musical compositions ought to be considered to be preliminary and the methods used for their calculation still require deeper investigations (this is emphasized particularly in the concluding part of the paper [85]).

Note also that the basic objective of statistical probability calculations, describing musical structure, in many cases does not at all consist of the determination of the entropy and redundancy. The fact is that the high degree of

redundancy to be found in high class music compositions allows us to give an entirely different, quite unexpected, application of statistical tables, which define the probabilities and conditional probabilities for different notes. In order to approach this application, recall the 'approximations of different orders' of an *English* sentence presented on pp. 178 and 181–184, i.e., the sequences of *English* letters, in which to a greater or lesser extent the intrinsic connection existing in the *English* language between adjoining letters was taken into account. It was seen that the farther we extended those relations, which are taken note of in the composition of our sentences, the 'more *English*' these sentences became, i.e., they tended to become closer in sound to the ordinary *English* language. However, it can obviously hardly be expected to obtain in this way completely meaningful expression, since there always exists some element of randomness in our sentences, which confuses their sense. Let us now attempt to apply the same methods to music. Here we shall obtain 'musical sentences' (i.e., the sequences of notes), all increasingly closer in their statistical structure to those sources which are used for the calculation of frequencies of different notes and their combinations. As in the case of 'models of English sentences', these new 'musical sentences' clearly shall not exactly repeat any of the sequences from a sample used for the calculation of frequencies. However, whereas in the case of language this situation makes our 'sentences' senseless, in the case of music it is expressly this which makes them remarkable; in fact, they represent new, original musical compositions!

Apparently, it is difficult to announce in advance the extent to which such 'models of musical melodies' may be of interest; it is also not clear to what extent statistical relations ought to be taken into account for obtaining compositions close 'in spirit' to the original material (i.e., for example, whether to imitate the compositions of a specific genre or of a particular author). It is, however, essential to note that by virtue of the appreciable redundancy of music we can arrive at sufficiently harmonious sounds via one of the earliest steps of the process described on p. 178 et seq. This was also shown convincingly in the earlier purely amateurish experiments of Pinkerton [145]. In these experiments, a note was taken of only the probabilities of individual notes and two-note combinations, which were furthermore strongly approximated. In particular, all the probabilities were rounded off to convenient fractions so that the choice of the next note could be made every time by drawing a card from a small collection of playing cards. Moreover, a simpler and cruder note-guessing procedure was suggested by Pinkerton which reduced all the probabilistic choices to a series of binary choices, so that at each step selection could be made simply by flipping a coin. Besides, by imposing auxiliary relations that assure the conservation of a definite rhythm of 'musical sentences' Pinkerton could obtain several new tunes which, according to the author's assertion, are sometimes not inferior to the original nursery tunes from the album used by him. The notation of one such 'randomly obtained' tune is given below (cf. [145], p. 84):



The redundancy of this tune can be calculated with comparative ease by starting from the statistical laws used for obtaining it; it is found to exceed 63%. In the words of Pinkerton, "this tune is highly monotonous, but nevertheless, less monotonous than some actual nursery tunes." Hence it can be inferred that in actual nursery tunes the redundancy is probably of the same order.

Similar attempts to obtain new melodies by means of experiments in which cards were drawn from an urn were carried out by F. and C. Attneave in relation to cowboy songs. Here also only the probabilities of individual notes and two-note combinations were considered (i.e., 'sentences' of the sort mentioned on p. 182 were constructed) and in addition it was required as well that a definite rhythm be preserved. The only difference from Pinkerton's work consisted of the fact that it turned out to be more convenient to compose the cowboy melodies from their 'end' by using the computed conditional probabilities of the notes *preceding* some given note. As shown at the London Conference on Information Theory, among several tens of 'random musical sentences' composed by Attneaves, two were found to be apt, which resemble the genuine cowboy melodies. The comparatively small percentage of success is naturally explained by the fact that only the simplest statistical regularities of the considered songs were taken into consideration.

The basic goal of Brooks *et al.* [80] was the same, namely to compose new melodies by means of 'random experiments'. In the given case only the 'draws of cards from card collections' were effected automatically by an electronic computer; operations of this type are found to be highly fruitful in many calculations using such computers (the so-called Monte-Carlo methods), and at present there exist well-developed methods for their automatic accomplishment. The immense potentialities of modern high-speed computers are demonstrated, in particular, by the fact that Brooks *et al.*, were able to compose all possible 'models of musical sentences' from 'first-order approximation sentences' in which only the relative frequencies of the appearance of individual notes (of the sort of the 'English sentence' mentioned on p. 181) were considered, and up to 'eighth-order approximations' inclusive, in which the frequencies of all possible sequences of eight notes were taken into account. For the composition of '*n*th order' sentences (where in different experiments *n* takes the value 1, 2, 3, 4, 5, 6, 7 or 8) a definite 'rhythmic scheme' was preassigned each time (relating to the distribution of durations of notes and rests), and then all notes were successively chosen 'at random' but in conformity with the computed frequencies of the different combinations of *n* notes. If subject to such choice, the given 'rhythmic scheme' was found to be unsatisfied, then the corresponding note was rejected and the computer automatically repeated the procedure

of 'random choice'; if 15 consecutive attempts resulted in 'rejected notes', then the computer was shut down and the composition of the entire series of notes was started afresh. In all nearly 600 'new hymns' were composed in this way (out of a total number of attempts of the order of 6,000); the high percentage of failures is explained by the fact that for some values of n (in particular, for $n = 5$ and $n = 7$) it turned out to be very difficult to satisfy the rhythmic scheme. Examples of melodies constructed with $n = 1, 2, 4, 6$ and 8 are listed below. For $n = 1$ and $n = 2$ the 'melodies' constructed contained many odd



combinations of notes and unnatural intervals; these 'melodies' are difficult to sing in spite of the presence of a rigid rhythmic scheme. For $n = 4$ and $n = 6$ they tend quite appreciably to sound like ordinary hymns. In the case of $n = 8$, the 'compositions' of the computer reduced to nonoriginal compilations: the rather lengthy parts of 'melodies' obtained coincide completely with fragments of one of the hymns and it is just occasionally (in places where two or more of the 37 hymns considered have the same groups of 7 notes) that a passage from one hymn to the other takes place (in particular, the fragment

written above is formed of a portion of three different hymns; the transition places are indicated by the brace appended below it). This position stems from the small volume of material utilized for the compilation of the frequency table, which naturally led to exceptionally high redundancy.† The fact is that many combinations of 8 notes did occur just once in the analyzed fragments of the hymns; hence for $n = 8$ many notes in succession were found to have been chosen from a single hymn.

Endeavours in an allied direction are also described by Olson and Belar [139], who make use of an analysis of the frequencies of individual notes, their pairs and triples in Foster's songs in order to evolve a special 'computer-composer' to compose (and then even play) quite simple musical compositions, reminiscent of Foster's melodies. Lately, experiments on the computer composition of artificial musical melodies by using the appropriate statistical analysis data have received a great impetus in a number of countries. For example, in the USA 'computer created' melodies are regularly played over radio and put on records or tapes, which are offered for sale. However, we shall not dwell here upon the indicated experiments, which are only indirectly related to a study of the information-theoretic characteristics of musical texts, but refer the interested reader to Pierce [17, Chap. 13] and Zaripov [177], who have considered all these experiments in great detail.

4.3.4. *Transmission of continuously varying messages. Television images*

Before we proceed further, we shall emphasize a fact that is of great importance for both theoretical and practical information transmission through communication channels. It is clear that spoken language or music differs principally from written language in the following respect: here the 'possible messages' are not sequences of symbols ('letters'), which can take a *finite* number of values, but are collections of sound vibrations, which can vary in a *continuous* manner. Hence, strictly speaking, it is necessary to consider that each sound can have infinitely many 'values'; but in that case all the formulas of our book become inapplicable. In the foregoing, we circumvented this difficulty by resorting to a decomposition of all sounds of the spoken language into a finite number of phonemes, and all musical sounds into a finite number of notes. But, is this legitimate?

For an answer to this question it is necessary that the decomposition invoked be understood in the true sense. The point consists in this that if we are interested in just the 'semantic' information contained in speech, then we cannot take notice of every variation of the speech sound if it does not obstruct

†Note that in any fragment, in which no N adjacent notes (or letters, or phonemes) are repeated, the entropy H_N is zero, i.e., the redundancy calculated with respect to H_N is unity. Hence reliable determination of the conditional entropy H_N for large N involves the use of a vast amount of statistical material.

our understanding of what is said and does not alter its meaning. Hence we can fully combine a majority of sounds that are similar among themselves if only the replacement of one of them by the other does not alter the meaning of what has been said. But a phoneme is also actually precisely such a collection of sounds close to each other and having the same meaning value (conversely, in speech the replacement of one phoneme by another can alter the meaning of a word; this property often forms the basis of the definition of a phoneme). Hence, clearly, when considering the problem of the 'semantic' information contained in speech, we ought to consider that the 'basic elements' of speech are not all sounds that are different among themselves (whose number is obviously infinite), but only a few 'intelligible sounds' having different meanings, i.e., phonemes. Exactly so is the case of music; if we are interested in just the information contained in the performed composition, but not in the interpretation of the composition by a certain performer, then it is necessary to identify all sounds that are expressed by the same sequence of note symbols, i.e., to consider only a finite number of different 'basic sounds' corresponding to a finite number of existing notes.

But one can pose an even broader problem. In particular, in the case of speech, besides 'semantic' information one can consider also the information contained in the intonation as well as the tone of the voice, and in the case of music one can be especially interested in the peculiarities of a given individual performance (the transmission of these peculiarities is a very important problem in communication engineering). The question is whether it is necessary in this case to consider that every sound can take an infinite set of values and hence can have an infinite entropy. A negative answer to this question has already been given on pp. 218-219, where we have deduced an evaluation of the entropy of spoken language with regard to different forms of 'unsemantic' information. We shall now undertake a more elaborate discussion to clarify this fact.

It is certainly true that the loudness of sound or the pitch of tone can be varied continuously, i.e., can take an infinite number of different values; moreover, in principle these values can replace each other as quickly as desired. However, our ear can distinguish only sounds that do not occur in extremely rapid succession; hence it can be considered that all sounds that we hear have a definite minimum duration. Moreover, we can discern only such sounds as differ in loudness and pitch by a bound not less than a certain definite finite value, and we cannot grasp a sound that is too high, or too low, or too soft, or too loud (loud sounds deafen us). Hence it follows that in fact only a finite number of scales of loudness and pitch of tone are distinguishable. By identifying on this basis all sounds, whose loudness and pitch of tone are determined to be within the range of one scale, we again arrive at our familiar sequence of signals, which can take only a *finite* number of different values.

The extremely general situation considered here is quite similar to the one

we were confronted with in the solution of Problem 22 in Section 2.3 (p. 80). There also we encountered the case of experiment β , having an infinite number of possible outcomes; however, it was found that for solving the problem experiment β can be replaced by a new experiment β_ϵ , obtained from β by identifying all its outcomes, which differ from each other by less than a small number ϵ . The entropy H_ϵ of β_ϵ (in contrast to the entropy of experiment β itself, H_ϵ is a *finite* quantity) is called the ϵ -entropy of β . In all problems concerning the transmission of messages, which are represented by *continuously varying quantities*, the ϵ -entropy occupies a very important place. In the transmission of such messages, a collection of all possible values of the signals to be transmitted is always partitioned into a finite number of scales ('cells' in the space of values) and all values within the range of one scale are identified among themselves (for instance, they are considered to coincide with the 'centre' of the corresponding cells). This operation of replacing a continuous message by a new message that takes only a finite number of possible values is called in communication engineering the *quantization* of a message. A quantized message always has a finite entropy (representing one of the variants of the ϵ -entropy of the original continuous message) that depends on the choice of the quantization method applied, but characterizes also the degree of uncertainty of the original continuous message. The latter circumstance decides us in favour of the possibility of using corresponding quantities in communication engineering.

An important class of such continuously varying messages is the *images* transmitted through television or phototelegraphic communication channels. It is easy to comprehend that principally we have here the same position as in the case of sound transmission — our eye is capable of distinguishing only a finite number of brightness grades of pictures and only those elements that are not too close to each other. Hence any picture can be transmitted 'through points', each of which is a signal taking only a finite number of values. In the case of phototelegraphy, in many cases we can consider that each 'elementary signal' (i.e., the smallest distinguishable element of a picture, the point) takes only one of two values, namely 'white' or 'black'; in black-and-white television it is necessary to take account of a considerable number (several tens) of grades of darkening ('brightness levels') for every element. In addition, phototelegraphic images are stationary, but on a television screen 25 still pictures are shown every second one after the other to create an effect of 'motion'. In both cases, however, no outcome of experiment α_0 , which consists of determining the value of a continuously varying image hue or brightness (varying from point to point, and in the case of television, varying in time also), is actually transmitted over a communication channel but rather the outcome of an altogether different 'quantized' experiment α_1 , which consists of determining the colour (black or white) or luminosity scale for a finite number of 'points'. This new experiment α_1 can have only a finite number of outcomes, and we can measure its entropy H (which is essentially a variant of the ϵ -entropy of the original experiment α_0).

The total number of elements ('points'), into which a picture has to be decomposed, is determined in the first place by the so-called 'resolving power' of the eye, i.e., its capacity to distinguish similar sections of a picture. In modern television, this number is usually of the order of several hundred thousands (in Russian telecommunication, a picture is decomposed into 400,000–500,000 elements, in American into approximately 200,000–300,000, in the transmission at certain French and Belgian television-centers into almost 1,000,000). It is easy to understand that for this reason the entropies of television images have vast magnitude. Thus, even if it is assumed that the human eye differentiates only 16 different 'brightness levels' (the value is evidently too low) and that a picture decomposes into altogether 200,000 elements, then the 'zero-order entropy' is found here to be $H_0 = \log 16^{200,000} = 800,000$ bits. The value of the true entropy H is obviously less, since a television picture has the significant redundancy $R = 1 - (H/H_0)$. Indeed, while calculating H_0 it has been assumed that the values of brightness at any pair of 'points' of a picture are independent of each other, whereas in fact the brightness usually varies very little in the passage to the adjacent elements of the same picture (or even a different but closely following one). The descriptive meaning of the redundancy R is that, among our $16^{200,000}$ possible combinations of brightness values at all points of a screen, the sensible combinations, which can be called 'pictures' form only a negligibly small part. An overwhelming majority of these combinations make up a completely disordered collection of points of different brightness, far removed from the 'subject' whatever it may be. On the other hand, the real 'degree of uncertainty' H of a television picture should obviously take note of only those combinations of brightness values that have at least some chance of being transmitted, and not all general combinations of brightness values.†

The determination of an exact value for the entropy H (or redundancy R) of a television picture demands a penetrating study of the statistical relations between the brightness of different screen points. This problem is quite involved and at present we have just a few relevant particular results. Schreiber [157] has measured, in particular, the values of the entropies H_0 , H_1 , H_2 and H_3 for a

†One should not merely think that the extreme scarcity of 'sensible pictures' automatically implies the redundancy R to be necessarily quite large. In fact, by assuming, say, that the human eye differentiates in all 10 different scales of brightness (so that the total number of possible brightness combinations is $10^{200,000}$) and that the 'sensible pictures' (which for simplicity are considered to be equally probable) form in all 0.00 . . . 01% (where 1997 zeros after a decimal point occur!) of all possible brightness combinations, it is easy to find that the redundancy R is close to $1 - [(200,000 - 2,000)/200,000] = 0.01 = 1\%$, i.e., it is extremely small (if the number of distinctive brightness scales were increased, then it would be still smaller). This apparently unexpected result is explained by the extremely slow variation of the function $\log n$ for large values of n , already mentioned on p. 208 (in connection with the evaluation of hieroglyphic writing) and on p. 218 (in relation to the estimation of 'unsemantic' information of speech).

number of television subjects of varying complexity, but he published the results for only two of them, of which the first (picture *A*, representing a landscape with trees and architectural structures) is the most complicated, and the second (picture *B*, representing a rather dark gallery with passers-by) is the most monochromatic in colour and contains the least details. Schreiber distinguished 64 different brightness levels of an element of a television picture; hence the entropy H_0 (related to one element, but not to the whole image) is found to be $H_0 = \log 64 = 6$ bits. Furthermore, with the aid of a special engineering device he calculated for both the considered pictures the relative frequencies (probabilities) p_1, p_2, \dots, p_{64} of all differentiable brightness levels and defined corresponding 'first-order entropy' by

$$H_1 = H(\alpha_1) = -p_1 \log p_1 - p_2 \log p_2 - \dots - p_{64} \log p_{64}$$

(note that a direct calculation of p_1, p_2, \dots, p_{64} can hardly be accomplished without the mediation of radio engineering when the total number of screen elements is of an order 200,000). The same engineering device was then applied for calculating the relative frequencies p_{ij} of the adjacent (horizontal) pairs of elements, in which the first (second) element has the i th (j th) brightness value, as well as the relative frequencies p_{ijk} of the adjacent (here also only horizontal) triples of elements, in which the first, second and third elements have, respectively, the i th, j th and k th brightness value (the numbers i, j and k run through all values from 1 to 64). These frequencies enabled him to determine the 'entropies of compound experiments':

$$H(\alpha_1 \alpha_2) = -p_{11} \log p_{11} - p_{12} \log p_{12} - \dots - p_{64,64} \log p_{64,64},$$

and

$$H(\alpha_1 \alpha_2 \alpha_3) = -p_{111} \log p_{111} - \dots - p_{64,64,64} \log p_{64,64,64},$$

and then also the conditional entropies:

$$H_2 = H_{\alpha_1}(\alpha_2) = H(\alpha_1 \alpha_2) - H(\alpha_1),$$

and

$$H_3 = H_{\alpha_1 \alpha_2}(\alpha_3) = H(\alpha_1 \alpha_2 \alpha_3) - H(\alpha_1 \alpha_2),$$

though H_3 was calculated only for picture *B*. The results obtained are tabulated below:

	H_0	H_1	H_2	H_3
Picture <i>A</i>	6	5.7	3.4	—
Picture <i>B</i>	6	4.3	1.9	1.5

From the table it is seen that the entropy H_1 does not differ very much from H_0 , it being appreciably larger for the picture A than for the picture B (this is obviously due to the greater monochromation of B in comparison to A). The conditional entropy H_2 (i.e., the average 'degree of uncertainty' of the brightness of a screen element when the brightness of the adjacent horizontal element is known) differs substantially from H_0 ; this also is remarkably lower for B than for A , which corresponds to the abundance of detail being less in B . The redundancy R , estimated with respect to H_2 [i.e., the difference $1 - (H_2/H_0)$] for A is 44% and 68% for B ; the real value of the redundancy can only be larger than this. As to the conditional entropy H_3 , when the brightness of two preceding elements of the same line is known, it differs comparatively less from H_2 (its corresponding redundancy value for B is 75%); hence we can conclude that by knowing the brightness of the closest elements we determine a very considerable part of the total redundancy.

The works of Lebedev and Piil [121] (see also [122]) and Limb [123] are also of a similar nature. In [121] and [122] some results are deduced from the calculations that are based on the use of statistical material slightly poorer than in [157] and a division of possible values of the brightness of an element of a television picture into 8, but not into 64 scales. These results include the evaluation of the entropies H_0 and H_1 and a number of conditional entropies H_2 , H_3 , and H_4 of a single element of the image for the following four television sport features: (A) fast running basketball players, (B) close-up of a spectator in the grandstand; (C) a panoramic view of spectators in the grandstand, and (D) fast running football players. Let us denote by the digits 1 and 2 the image elements adjacent to the given element 0 in the horizontal and vertical direction, by 3 the adjacent diagonal element, by 4 the same element as the given one but considered in the preceding television transmission frame, by 5 an element at the same line adjacent to element 1 and, finally, by 6 the same element in the frame, preceding the one which contains element 4 (see Fig. 16a). We set up in upper parentheses of conditional entropy notation the image element numbers,

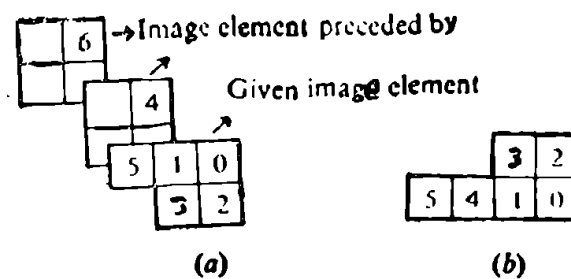


Fig. 16.

whose brightness level is assumed to be known. In such a case, following [121] (see also [122]) the values of various entropies (in bits) can be presented in the form of the accompanying table.

	H_0	H_1	$H_2^{(1)}$	$H_2^{(2)}$	$H_2^{(4)}$	$H_2^{(6)}$
(A)	3	1.96	0.69	0.98	—	1.77
(B)	3	1.95	0.36	0.39	—	—
(C)	3	2.78	1.34	1.95	2.78	—
(D)	3	2.45	—	—	2.00	2.08

	$H_3^{(1,5)}$	$H_3^{(4,6)}$	$H_3^{(1,2)}$	$H_4^{(1,2,3)}$	$H_4^{(1,2,4)}$
(A)	0.68	—	0.56	—	—
(B)	0.35	—	0.27	0.26	—
(C)	—	—	1.22	1.18	1.19
(D)	—	1.83	—	—	—

(a dash in this table signifies that the corresponding entropy has not been computed). The following four parts (each containing 5,000 individual elements) of two television images are analyzed in [123]: (A) the earth surface of an average setting covered with grass and bushes; (B) a part of the same landscape, adjacent and similar to (A); (C) a part of sky with clouds of comparatively uniform light hue; (D) close-up of a large grassy area with bushes. The image elements are divided into 16 brightness levels; for calculating the conditional entropies of an image element with number 0, the data related to the elements 1, 2, 3, 4 and 5 of the same and preceding lines of the same frame (see Fig. 16b) are used. The results obtained in [123] are listed in the accompanying table.

	H_0	H_1	$H_2^{(2)}$	$H_2^{(1)}$	$H_3^{(1,2)}$	$H_3^{(1,4)}$	$H_4^{(1,2,3)}$	$H_4^{(1,2,4)}$
(A)	4	2.85	2.24	2.38	1.82	2.10	1.46	1.47
(B)	4	2.51	1.99	1.96	1.66	1.66	1.15	1.28
(C)	4	1.32	1.04	0.99	0.94	0.97	0.90	0.92
(D)	4	3.72	2.70	3.10	2.01	2.23	0.87	0.86
(A) and (B)	4	2.90	—	2.27	—	2.03	—	1.54
(C) and (D)	4	3.29	—	2.17	—	1.65	—	0.91
(A), (B), (C) and (D)	4	3.52	—	2.31	—	2.00	—	1.39

The data contained in [121]—[123] are qualitatively close to the results in [157] (a quantitative comparison is difficult here because of the differences in the number of quantization levels used, affecting the numerical values of the entropies) but are considerably more complete. In particular, Schreiber's conclusion (related to comparative monochromatic and detail-starved image *B*) to the effect that when a preceding image element is known, a further knowledge of any other elements alters but little the degree of uncertainty (i.e., the entropy) of a given element of a television image is in excellent agreement with the data related to the monochromatic and detail-starved images of a closeup person's face (image *B*) of [121], [122] and a cloudy sky (image *C* of [123]). It may, however, be noted that according to the data deduced in [122] the stated conclusion does not fare badly even for all other investigated images (including also the most 'heterochromatic' image *C*), while the results of [123] related to the images (*A*), (*B*) and (*D*), do not corroborate it. An analysis of Limb's data permits us to infer also that the use of probabilities (i.e., frequencies) calculated for a large and quite inhomogeneous image (whose model can be represented by the union of the heterogeneous parts of (*A*), (*B*), (*C*) and (*D*) in two different frames) leads to just a small increase in the values of the conditional entropies (when the brightness values of one, two or three preceding elements are known) in comparison to the values of the conditional entropies calculated for the parts of the image taken separately. Furthermore, the results of [121], [122] related to the conditional entropies, when brightness values of the same image element at one or two preceding frames are known, show that for the rapidly changing images under consideration these conditional entropies exceed appreciably the conditional entropy, given the brightness of the preceding (along the line) element of the same frame. Hence, by considering the relation between the brightness values on the succeeding frames in television transmission it is not possible to adduce here the considerable increase of redundancy determined from the analysis of the brightness distribution in one frame. The preceding conclusion apparently may not be valid for television subjects, for which the image varies less over time; however, reliable quantitative data, related to such cases, are still lacking (some estimates of time relations, based on indirect arguments, may be found in [132]). The total redundancy of television images by the data in [123] both in the case of an image rich in detail (a 'close-up of vegetation') and in that of a detail-starved image ('sky') is found to be not less than 80% (but for a 'medium' image (*A*) or (*B*) it turns out for some obscure reason to be not so high, although it is nevertheless not less than 65%). At the same time the results in [121], [122] lead to the conclusion that for a detail-starved image ('of the face of a person') the redundancy is not less than 90%, and for a detail-affluent image ('of many spectators'), it is not less than 60%. Note that, the values of redundancy in [121]—[123], larger than those found by Schreiber [157], can be naturally explained by a cruder division of the brightness levels. On the other hand, the divergences in the conclusions due to Lebedev and Limb regarding the differences

of redundancy between 'monochromatic' and 'heterochromatic' images are related to the disagreement remarked above in the results of these authors on the rate of decrease of entropy values in the sequence H_0, H_1, H_2, H_3, H_4 for all images not too poor in details. (The reasons for this disagreement are not yet clear, but on the whole the results in [121], [122] seem to be more plausible than those in [123].)

It is clear that calculations of the sort set forth in [121]—[123] and [157] cannot be used for the determination of relations which affect the redundancy of an image, between many elements. In fact, even in the case of entropy H_4 , the number of different combinations of brightness values at four points already turns out to be vast (recall that a comparatively crude division into brightness levels is applied in the works [121]—[123]), and with a further increase in the order of the conditional entropies, this number increases enormously which makes calculation intractable. Hence it is worthwhile to draw attention to a few efforts that were made to estimate the image element entropy and redundancy via Shannon's 'guessing experiment method' (or some other method which does not involve the calculation of the frequencies of groups of many image elements).

Apparently, the first, still imperfect, attempt in this direction was made by Parks [141]. He tried to apply guessing-experiment method (and also a cruder method of restoring the entire picture when only a part of it is exposed) to the approximate estimation of the redundancy of three very different half-tone (i.e., black-and-white) pictures. Of these, the first (a close-up portrait of a sailor) contained the least details in comparison to the rest. The second picture (of a girl with a flower lying on a rug) ranged between the other two and the third one (a reproduction of an abstract painting) was the most variegated of all. All pictures were divided by Parks into about 1,500 square elements, and the average gray-shade (i.e., the degree of blackness) was determined for every element. Then all gray-shades of different elements were divided into eight levels in case of the first and the third pictures and into six levels in case of the second picture. Parks covered all pictures with an array of square tiles corresponding to all the picture elements and asked a number of subjects (selected from the undergraduate university students in fine arts, who were unfamiliar with the picture) to perform guesses. Expressly, every subject was asked to remove any tile of his choice and then guess the blackness levels of all the remaining picture elements in any order he desired. After every guess the corresponding tile was removed and the subject could use the knowledge of the true shade of the element in guessing about the next tile. Parks does not describe in detail the experiments that were performed and gives only the estimates finally arrived at, which are apparently rather crude. According to his results, the redundancy is not less than 75%, 66% and 40% for the first, second and third pictures, respectively.

A second more simple but a still cruder method of redundancy estimation was based on the following guessing experiment. Beginning with the picture fully covered with tiles, a constant percentage of tiles was randomly removed

and the subject was asked to describe the picture. Then the experiment was repeated with a higher percentage of removed tiles until the answer was considered to be fully correct. This method may clearly give only the strongly underestimated values of redundancy, which enabled Parks to draw the conclusion that the redundancy of the first and second pictures is apparently significantly above 75% and 50%, respectively (in case of the third picture this method did not work satisfactorily).

Later Tsannes and his students at Tufts University [168] made a more thorough attempt to apply the guessing-experiment method due to Shannon (described on p. 188 et seq.) for an evaluation of the conditional entropy of an image element with regard to higher order spatial relations between the elements. Tsannes chose as original material 20 photographs of a section of lunar landscape, each of which was represented in the form of a collection of $50 \times 50 = 2,500$ individual elements, taking one of eight possible values according as its 'degree of blackness' (i.e., the levels of blackness). These photographs were further divided into four groups of photographs of a similar nature. One of the photographs (together with its numerical form, representing a quadratic table of 2,500 numbers from 0 to 7) was given to a guessing person (a senior student at the university), who studied it attentively. (The 'familiarity with the picture' attained in this way is obviously very poor in comparison with the knowledge of the structure of the mother-tongue inherent in every literate person, which is used in the guessing experiments related to written text, but this is inevitable). After this study, the same person proceeded to guess successively the elements of another photograph in the same group. In the course of guessing, movement was permitted in any direction after each already guessed element; to each conjecture the answer 'yes' or 'no' was given, which was considered to contain one bit of information (in fact, it often contained considerably less information since both possible answers are not at all equally probable). Thus, the average number of questions per element of the image provided a rough estimate from the above (i.e., a strongly overstated estimate) of the average entropy of one element of the image. In the two guessing experiments described in [168], this average estimate turned out to be roughly 1.8 bits in one case and 1.3 bits in the second; the authors remarked that an expert in lunar landscapes, by dint of his prior practice in this field, would probably obtain a remarkably better result (i.e., a lower bound on the entropy). In any case, both estimates so obtained are found to be appreciably lower than the value $H_0 = 3$ bits; the true entropy H is obviously significantly lower than these estimates. If, following Shannon's proposition mentioned in the footnote on p. 188, use is made only of the result of the more successful of the two guessing persons, then the corresponding lower bound on the redundancy of the lunar surface image comes close to 60%.

The appearance of colour television has also given rise to the need to estimate the information contained in the colour of the image. By way of a rough guide, the pioneer calculations in this direction have shown that for colour

television images, matching in quality the colour illustrations in magazines, the information, in order of magnitude, compares to double the information contained in the corresponding black-and-white images (see [132]).

4.3.5. Phototelegrams

Let us now take up the data concerning a *phototelegram*. Here, the general principle of image transmission is close to the telecommunication principle: the image is split into smallest squares ('screened elements'), after which the information on the colour of each such element (whether it be black or white) is transmitted over the channels. Thus, compared to black-and-white television the images now considered are simpler: for them there are no brightness grades (i.e., degree of blackness) and the colour can take just two values. It is natural that the maximum information (i.e., the entropy H_0) contained in the knowledge of colour per element equals $H_0 = \log 2 = 1$ bit; this information is attained when black and white elements occur with the same frequency and the colour of each element is independent of that of all the rest. In reality, the two colours usually occur with different frequencies (the number of white elements as a rule considerably exceeds the number of black ones) and between the colours of individual elements there is a noticeable dependence; hence the true value of the entropy of one element of a phototelegram is appreciably less than 1 bit. The task, therefore, is to determine its value.

It can be calculated that, in the transmission of the printed text from an ordinary book or magazine by phototelegram, the relative frequency p_0 of white elements is close to 0.8, and the frequency p_1 of black elements is close to 0.2. Hence the entropy H_1 is given by

$$H_1 \approx -0.2 \log 0.2 - 0.8 \log 0.8 \approx 0.72 \text{ bit},$$

which corresponds to the redundancy $R = 1 - (0.72/1) = 0.28 = 28\%$. However, this value of redundancy is grossly understated since it takes no note of the dependence between the colours of adjoining elements. Unfortunately, a precise quantitative estimation of this dependence (stretching to a large number of adjoining elements) is highly involved; hence even approximate methods are of interest for evaluating the entropy H_∞ and redundancy R .

One of the earliest attempts, quite a sketchy one, to estimate the entropy $H_\infty = H$ of a phototelegraphic message is traced to the work of Deutsch [89]. In this work he analyzed a small fragment of an *English* text (a few lines long) printed in comparatively large letters. Unfortunately, a text written on paper is not in the least easy to divide directly into very small 'screen elements' employed in a phototelegram and in case of such division the given fragment turns out to consist of a vast number of elements, which make the arithmetic calculation of the frequencies of different combinations exceptionally tedious. Hence, Deutsch resorted to a partition of the given text into comparatively larger

squares, each consisting of many screen elements. He classified such squares as white or black according to which colour predominated in a square (e.g., if more than 50% of the area of the square is found to be white, then the whole square is considered to be white; otherwise, black). It is natural that in such a case $H_0 = \log 2 = 1$ bit for a 'square' and a screen element as well. Furthermore, Deutsch calculated the conditional entropies H_1 , H_2 and H_3 for vertical 'blocks' consisting of several adjoining squares (for horizontal 'blocks' only the entropy H_2 was calculated, which turned out to be slightly larger than the corresponding value for the vertical 'blocks'). The entropy H_1 was found to be 0.67 bit, which conformed to the redundancy R being 33%; the entropy H_3 had already a value of 0.57 bit, i.e., it corresponded to the redundancy $R = 43\%$.† By means of some indirect arguments, it was also shown in [89] that the entropy of one 'square' must in fact be considerably less than 0.5 bit, so that here the redundancy R does significantly exceed 50%. Note, however, that all these figures do not merit any particularly great reliability, since the partition of the text employed in [89] into comparatively large squares distorts considerably its statistical structure.

The German scientist Kayser [112] carried out a study of this sort in considerably greater depth. He decomposed the typewritten text into much smaller

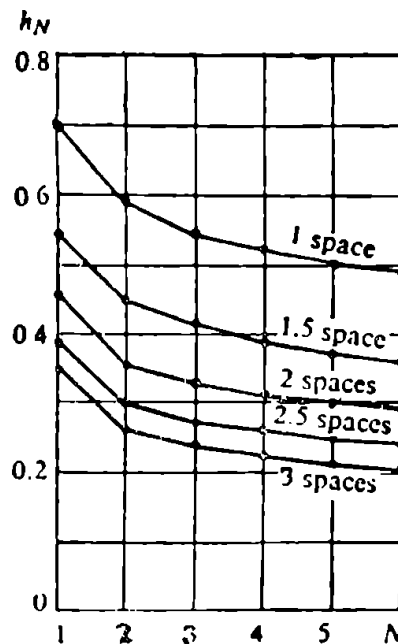


Fig. 17.

†For vertical blocks the entropy $H^{(N)}$ of a block of N adjoining elements for $N = 1, 2, 3$ and 7 was calculated. It is interesting that the ratio $H^{(N)}/N$ for $N = 7$ was found equal altogether to 0.58 bit, i.e., slightly larger even than H_3 . This fact clearly shows to what extent the sequence of quantities $h_N = H^{(N)}/N$, $N = 1, 2, 3, \dots$ tends more slowly to H_∞ than the sequence H_N (see footnote on p. 184),

squares with a side of 0.2 mm (one typed page here was found to have been divided into roughly a million individual elements). In order to make the calculations possible with so large a statistical ensemble, Kayser constructed a special measuring apparatus to separate automatically the succeeding 'blocks' of a small number of N adjoining elements and register the number of blocks of distinct composition. This apparatus was then applied to blocks in different directions (horizontal, vertical and positioned at an angle to the typed text), and all enumeration results were found to vary only slightly with changes of direction. Starting from here, Kayser confined the data analysis mainly to horizontal blocks. In relation to such blocks he investigated the dependence of the specific entropies $h_N = H^{(N)}/N$, with $N = 1, 2, 3, 4, 5$ and 6 , on the following factors: (a) the extent of 'boldness' (i.e., the thickness of the letters) of the text, (b) the distance between the lines, and (c) the size of the typescript (i.e., the degree of magnification of the typescript copy). The results obtained by him with respect to the text of standard 'boldness' and size and five different distances between the lines (from the densest typescript 'through a single space' up to the least dense typescript 'through three spaces') are shown in Fig. 17. From this it is seen that the redundancy of the most densely typed text (though normal in all other respects) certainly exceeds 50%, whereas for the least dense typescript it is not less than 80% (but, apparently, these figures are sharply understated, since h_6 is a very rough estimate of the quantity H_∞). In the case of thinly typed text all entropies naturally turn out to be smaller, and the redundancies larger than those for a standard text; particularly, the value of $h_1 = H_1$ is appreciably reduced; however, for increasing N the values of h_N for thin print tend progressively to the values for ordinary print. Conversely, all entropies for very 'heavily' set text are found to be larger than those for normal text, the greatest difference being again observed for $N = 1$, and the smallest for $N = 6$. For homothetic magnification of a typed text the values of $h_1 = H_1$ are not affected (since the fractions of white and black elements are not altered), but in this case the statistical relations between the adjoining elements increase, and hence all entropies h_N with $N > 1$ decrease but the redundancies increase. In relation to the values of h_N with $N > 6$, only some quite rough estimates are deduced in [112], according to which, say, for a single-spaced standard typed text, $h_{12} \approx 0.40\text{--}0.45$ bit.

It is clear that the quantities h_N for small N by no means characterize the complete redundancy of a typed text brought about by all statistical relations existing in such text. This is seen, in particular, from the fact that by applying quite a different method Kayser obtained results that differ sharply from those described above. The measuring instrument he constructed surely could not properly recognize the fact that all black elements in its field of vision are portions of 26 letters of a well-defined form. Hence Kayser did further work to determine what is the smallest segment of a square closely covering a letter, by the sight of which a literate person is able to guess what letter it is. The

experiments undertaken with this objective showed that if for each letter its most characteristic part is selected, then it suffices to show only about 15% of the area of the square. Hence it can be inferred that the redundancy of a two-dimensional figure of an individual letter (and, consequently, also of a very closely typed text) on an average is close to 85% (the blank spaces between the letters, words and lines in a printed text in general can be considered as entirely redundant, however). In addition, we must note that only a part of an isolated letter was shown; however, if the entire text preceding this letter were known in advance, then quite often the letter can be guessed correctly even without looking at any part of it. Hence it is clear that the size of the part needed to guess one text letter would be on an average appreciably less than 15%. Starting from the data of [118], which is mentioned on p. 195, Kayser concluded that knowledge of the preceding letters of a typed *German* text must decrease the limiting amount of uncertainty (i.e., H_∞) by approximately a factor of three. Hence he arrived at the result that the true redundancy of a closely typed text is obviously close to 95%. This redundancy estimate evidently makes allowance for the highly complex statistical relations covering simultaneously many 'screen elements', generated by both the ways of letter writing and grammar and structure of the language; their employment in phototelegraphic engineering is, of course, still a remote possibility.

In the following we shall no longer take note of the semantic and grammatical properties of phototelegraphic texts, and instead consider only the statistical regularities in the mere interchange of black and white screen elements. In this case a fairly good estimate of the entropy H of one screen element can be obtained by representing each line of a phototelegram in the form of a sequence of alternating white and black sections of different lengths. By calculating the relative frequencies of the appearance of all such sections the corresponding 'first-order entropy' $H_1^{(\text{section})}$ can be calculated; here the ratio $H_1^{(\text{section})}/w$, where w is the average number of elements in one section, is surely greater than the true value of the entropy H of one element (see the discussion on p. 187). By means of this method, Michel [130] showed that in the transmission of a densely typed ('single spaced') text, in large type, the entropy H is smaller than 0.3 bit, i.e., the redundancy R exceeds 70%; a similar conclusion is also obtained in [112] by using the same method. A more detailed investigation of this sort has been carried out by Garmash and Kirillov [103] on the basis of quite extensive statistical material for *Russian* printed book or magazine text. These authors calculated not only the frequencies of monochromatic sections of various lengths, but also the frequencies of all possible pairs of such sections and determined from this data the first-order section entropy $H_1^{(\text{section})}$ and second-order entropy $H_2^{(\text{section})}$. By calculating the ratio $H_1^{(\text{section})}/w$, they determined that in the transmission of printed text $H \leq 0.33$ bit, i.e., $R \geq 67\%$; the inequality $H \leq H_2^{(\text{section})}/w$ allowed them to refine further this estimate and show that

$H \leq 0.28$ bit and, correspondingly, $R \geq 1 - 0.28 = 72\%$.

Another method of estimating the entropy H and the redundancy R for a phototelegram is due to Vasiliev [171] and Frolushkin [99]. It is clear that an exact calculation of the entropy $H^{(N)}$ of an experiment, consisting of determining the colours of N successive screen elements, for large N , is highly involved because of the fact that the total number 2^N of outcomes of this experiment is extremely large. Hence we divide the corresponding 2^N outcomes into some n groups containing respectively M_1, M_2, \dots, M_n outcomes (where $M_1 + M_2 + \dots + M_n = 2^N$) and we determine only the probabilities q_1, q_2, \dots, q_n of the successive N elements belonging to the 1st, 2nd, \dots , n th group. Assume now that within each group all outcomes are equally probable (the nonfulfilment of this restriction can only *decrease* the entropy $H^{(N)}$!), and determine the value of $H^{(N)}$ subject to this assumption. In this case, the outcomes belonging to the i th group (where i can take the values $1, 2, \dots, n$) contribute M_i identical terms $-(q_i/M_i) \log (q_i/M_i)$ to the expression for $H^{(N)}$. This implies that

$$H^{(N)} \leq -q_1 \log \frac{q_1}{M_1} - q_2 \log \frac{q_2}{M_2} - \dots - q_n \log \frac{q_n}{M_n} \quad (*)$$

(the use of the \leq sign is connected to the fact that our calculations yield in general an exaggerated value of $H^{(N)}$). Similarly, by assuming that one of the outcomes of i th group has probability 1 and all the rest have probability 0, i.e., they are impossible (the nonfulfilment of this restriction can only *increase* the entropy $H^{(N)}$!), we obtain

$$H^{(N)} \geq -q_1 \log q_1 - q_2 \log q_2 - \dots - q_n \log q_n. \quad (**)$$

Vasiliev [171] started from the fact that in the transmission of printed text a quite significant part of the redundancy is related to the high frequencies of comparatively long sections of N white elements (which arise because of the presence of interline spaces and margins). In agreement with this, his first group of outcomes is formed from a single outcome, the one in which all N elements are white; the remaining $2^N - 1$ outcomes make up the second group. In this connection, formulae (*) and (**) yield

$$-q \log q - (1 - q) \log \frac{(1 - q)}{2^N - 1} \geq H^{(N)} \geq -q \log q - (1 - q) \log (1 - q),$$

where q is the probability of a 'white' block of N screen elements. If it is further noted that for large N the expression $2^N - 1$ is almost the same as 2^N , so that $\log (2^N - 1)$ can be replaced by $\log 2^N = N$, then

$$\begin{aligned} \frac{-q \log q - (1 - q) \log (1 - q)}{N} + (1 - q) &\geq h_N \\ &\geq \frac{-q \log q - (1 - q) \log (1 - q)}{N}, \end{aligned}$$

where $h_N = H^{(N)}/N$ is the approximate value of the 'specific entropy' of one screen element. In order to obtain a satisfactory estimate of $H = H_\infty = \lim_{N \rightarrow \infty} h_N$

it is necessary to take N of the order of ten or several tens; in this, q for newspaper text turns out close to 0.5 (or even more), and for typed text set in the ordinary way ('double spaced') close to 0.7 (or more). It is hence clear that in the transmission of newspaper text $H \leq (1/10) + 0.5 = 0.6$ and $R \geq 1 - 0.6 = 40\%$; in the transmission of ordinarily typed text

$$H \leq \frac{-0.3 \log 0.3 - 0.7 \log 0.7}{10} + 0.3 \approx 0.39 \text{ and } R \geq 1 - 0.39 = 61\%.$$

The value of such a comparatively rough estimate of the entropy H lies in the fact that here it is easy to specify a concrete coding method, which permits the transmission to be conducted at the rate

$$v = \frac{C}{H} = \frac{NC}{-q \log q - (1 - q) \log (1 - q) + N(1 - q)}$$

(per screen element/unit time), where C is the capacity of the communication channel being used (see [171]).

In [99], all blocks of N screen elements are partitioned into a large number of groups, characterized by definite values of 'saturation' and 'mesh'. By 'saturation' is understood here just the total number of black elements in a block (so that for a block of N elements the 'saturation' can take $N + 1$ values: 0, 1, 2, ..., N), and by 'mesh' the number of monochromatic sections into which a given block is partitioned (the 'mesh' of a block of N elements can equal 1, 2, 3, ... or N , i.e., can have N distinct values). The calculation of the values of 'saturation' and 'mesh' of individual blocks was carried out automatically by means of a convenient special device constructed by Frolushkin. The value of N taken in [99] is 100, i.e., the quantity $H^{(100)}$ is evaluated and the entropy H of one element is equated to $h_{100} = H^{(100)}/100$. In connection with such a choice of N the measuring circuit is provided with a device, which automatically switches the circuit on for the time interval, corresponding to the transmission of 100 screen elements of a phototelegram through a channel; after that the circuit is switched off, the values of 'saturation' and 'mesh' are registered and it is only after this that another section of the phototelegram is fed into the circuit.

Phototelegrams with handwritten, typed and printed (newspaper) texts were analyzed separately, where in all cases the phototelegrams were filled with the densest possible text, as is customary in real transmission. Each of the three types of texts was represented by 10 extracts and from every extract 400 distinct blocks of 100 elements were selected. From the data obtained the frequencies (approximate values of probabilities) of different values of 'saturation' and 'mesh'

were obtained, and also the frequencies of different combinations of the values of 'saturation' and those of 'mesh'. By calculating further the number $M_n^{(\text{sat.})}$ of blocks having the given 'saturation' n , the number $M_m^{(\text{mesh})}$ of different blocks having a given 'mesh' m and, finally, the number $M_{n,m}$ of blocks having simultaneously the 'saturation' n and 'mesh' m (all these numbers can be determined by means of simple combinatorial arguments†), and by using the formula (*) (p. 242) we obtain three different estimates of the entropy H (and, consequently, of the redundancy $R = 1 - (H/H_0)$). It is clear that all these estimates slightly overstate the value of H (and understate the value of R), though the third of them (corresponding to the division into the greater number of groups), in principle, ought to be more precise than its two predecessors.

Estimates of the values of H and R for the three types of text obtained as a result of these investigations are listed in the accompanying table. It is seen

	<i>Evaluation by the data on 'saturation'</i>		<i>Evaluation by the data on 'mesh'</i>	
	H (in bits)	R	\bar{H} (in bits)	R
Handwritten text	0.37	63%	0.22	78%
Typed text	0.53	47%	0.30	70%
Newspaper text	0.43	57%	0.34	66%
Average	0.44	56%	0.29	71%

that the estimate of H obtained from the data on 'saturation' is found to be appreciably cruder than the estimate conforming to the data on 'mesh'. Hence it can be inferred that the assumption of equi-probability of all different blocks with the same 'mesh' is in better agreement with the facts than that of all blocks with the same 'saturation'. In other words, the blocks with the same 'mesh' form a more homogeneous group than do those with the same 'saturation'.

The evaluation of the entropy H from the data on the probabilities of all possible combinations of 'saturation' and 'mesh' demands a considerable increase in the volume of the subject material. In fact, it is easy to calculate that for blocks of 100 elements it is possible to form in all nearly 5000 (precisely 5001) such distinct combinations. Consequently, an entire set of all possible different blocks (containing $2^{100} > 10^{30}$ blocks, i.e., the number of blocks expressed by

†It is easy to show that in the general case of blocks of N elements

$$M_n^{(\text{sat.})} = \binom{N}{n} = \frac{N!}{n!(N-n)!} \quad \text{and} \quad M_m^{(\text{mesh})} = 2 \binom{N-1}{m-1} = \frac{2(N-1)!}{(m-1)!(N-m)!}$$

(the latter formula follows from the fact that in this case the $m-1$ 'boundaries' between different monochromatic sections can be chosen in $\binom{N-1}{m-1}$ different ways, and after this the first monochromatic section can be chosen either as black or white by an arbitrary rule). As to the number $M_{n,m}$ it is given by a more complex formula, which we shall not deduce here.

a 31-digit number!) is split up here into 5001 individual groups. It is clear that the probabilities of all these groups are by no means possible to evaluate from the data on the frequencies obtained during the analysis of $400 \times 10 = 4000$ different blocks. Hence, the third estimate of the entropy is given in [99] only for 'average *Russian* text' (on the basis of the data on the frequencies of individual groups in the entire collection of analyzed blocks irrespective of the text from which they are extracted). This estimate, obtained by using formulae (*) and (**), has the form

$$0.23 \geq H \geq 0.06, \text{ i.e., } 77\% \leq R \leq 94\%.$$

Here the true values of the entropy H and redundancy R apparently ought to lie somewhere between the stated limits.

During our discussions on phototelegrams, we have so far considered only cases of the transmission of text material (handwritten, typed or printed) through a phototelegram channel. However, phototelegrams can also be used for the transmission of a number of different types of white-and-black messages, and for a number of them the values of the average entropy (for one screen element) and the redundancy may turn out to be quite different from that in a literal text. Thus, for instance, it is clear that in the case of a drawing the redundancy would be expected to be appreciably higher than in the case of a text (in the first place due to the fact that in a drawing 'black' occupies a much smaller place than in a sheet of literal text). This conclusion has already been verified by earlier (though highly crude and expressly appreciably overstated) estimates of the entropy H for drawings obtained (on the basis of data on the length distribution of monochromatic sections) by Michel in his work [130] mentioned above. According to the estimates due to Michel, in the case of intricate radio-circuit diagrams which include a series of inscriptions it can be confidently stated that $H \leq 0.12$ bit, i.e., $R \geq 88\%$, while for a simple drawing the entropy H can turn out to be less by even more than half (i.e., the redundancy exceeds 95%). A more accurate (but also considerably more complex) method for an approximate evaluation of the entropy and redundancy of simple drawings (consisting of a number of continuous lines) was set forth by Foy [95]. In the case of a model example analyzed in [95], the calculation of just the deviation of the relative frequency p_1 of black elements from $\frac{1}{2}$ led to the estimate $H \leq 0.08$ bit, $R \geq 92\%$ (here the value of p_1 is close to 0.01), whereas the employment of the more accurate method due to the authors permits one to obtain the following result: $H \leq 0.015$ bit, $R \geq 98.5\%$. As to the pictures and photographs to be transmitted through a phototelegram, these types of messages in fact differ little from black-and-white television pictures; hence we need not dwell exclusively on the data of their entropy and redundancy, and instead refer the reader to the preceding sub-section of this chapter.

4.3.6. *Capacity of real communication channels*

Let us now discuss briefly the question of the practical fruitfulness of the estimates of entropy and information of various messages in communication engineering. The role of entropy in the theory of message transmission is defined by the fundamental theorem of Section 4.2 (pp. 172–173). According to this theorem, the maximum value of the transmission rate v attainable over a communication channel is defined by the formula

$$v = \frac{C}{H} \text{ element/unit time,}$$

where H is the entropy of one element of a message (no matter whether it is a letter, phoneme, note, element of a teleimage, or screen element of a phototelegram), and C is the channel capacity. Hence, in order to find the limiting transmission rate it is necessary to know not only the entropy H , whose determination for different cases has been dealt in the preceding sub-sections, but also the capacity C . The question arises as to how to determine such capacity.

In section 4.2 it is seen that

$$C = L \log m,$$

where L denotes the number of elementary signals that can be transmitted through a channel in unit time and m denotes the total number of distinct signals to be used. In practice, the number m is often chosen with the condition that for the corresponding communication channel it is possible to set up a sufficiently simple and inexpensive transmitting and receiving device. Thus, for example, most often in all two elementary signals are taken (ordinarily, on and off current). This is due to the fact that the problem of distinguishing two such signals at the receiving end is technically most straightforward and corresponding receiving devices are most economical and reliable. However, for those cases in which it is required to transmit as many messages as possible within unit time, it is natural to ignore considerations of simplicity and economy of channel circuit and strive to increase to the maximum the values of L and m . And, at first glance the opportunities offered here seem to be completely unlimited: usually the signals transmitted over a communication channel may vary continuously, so that, as is apparent, it is possible to choose them as short in length and as slightly different from each other as desired. But this implies that L and m can be made as large as desired and, consequently, the capacity of any channel, transmitting continuous signals, is factually unbounded. The question arises as to what role is played in such case by larger or smaller values of the entropy H .

In reality, however, the arguments set out here are not true: any communication channel, transmitting continuous signals also has a strictly limited capacity. In the first place, the value of a transmitted signal can never be changed

instantly—for this a definite time is always required. In practice, in a communication channel being used the minimum time required for a noticeable alteration of a signal is strictly regulated by the engineering characteristics of the channel itself. This leads to the fact that for every channel only values of a signal at the time points divided into a definite minimum time interval τ_0 can be chosen more or less arbitrarily: after these values are chosen, all values of the signal are defined uniquely in the intervening instants of time. In other words, the maximal number $L = 1/\tau_0$ of distinct elementary signals that can be transmitted through a communication channel in unit time is a fixed characteristic, which cannot be altered without introducing changes into the channel itself. This position, which plays a central role in all applications of information theory to the problem of the transmission of continuous signals, was stated clearly even before the origin of modern information theory (in 1933) in a report by Kotel'nikov. The main result of Kotel'nikov's work which was also obtained independently by Shannon in [21] and [158] permits to express the number L in terms of the usual engineering characteristic of a communication channel (in terms of the so-called 'transmission band width'). The expression obtained shows, (say) in the case of radio-communication, that the replacement of a channel with the object of increasing the values of L may not bring an advantage since it makes the operation of other radio channels impossible, driving the transmission over close wavelengths (see, for example, [24], [25] or [115]).

But suppose that only the number m can be chosen as large as desired, then this suffices to attain as large a capacity C as desired. Unfortunately, even this assumption is not true. At the outset, we cannot use signals of arbitrarily high intensity since for this we have to utilize vast power for their production. There exists a strictly definite average power P for the signals to be transmitted, defined uniquely by the energy source of our communication channel. In addition, we also cannot distinguish signals whose values are too close to each other. We confronted this situation on pp. 228-230 where the maximal degree of closeness under which signals could be still distinguished, was determined purely by physiological factors ('resolving power' of the eye or ear). In the case of artificial communication channels, reception is effected by a special device, and at the price of modifying and further raising the cost of this device, its resolving power can be made practically as high as desired, i.e., a device that distinguishes between even extremely close signals can be produced. But there is one more factor that obstructs the discernment of close signals, noise. The fact is that in every communication channel there exist disturbances which can by no means be eliminated; these disturbances distort the value of the transmitted signal. In the case of electro-communication, for instance, these disturbances can be produced by small oscillations of the load in the network, by the electrical field of adjacent circuits and neighbouring electrical machinery, or even just by the random 'thermal' motion of electrons in the conductors (this motion depends on the conductor temperature and is completely similar to the chaotic motion

of gas molecules). In the case of radio-communication they can originate in lightning discharge in the atmosphere or electrical discharges created by industrial or transport facilities (say, by the sparking of the arc from a nearby passing tram) and so on. If we denote by W the average power of these disturbances (i.e., the power of those distortions to which our signals are subjected in the process of transmission), then those signals, between which the variation in power is much less than W , are impossible to distinguish by any device at the receiving end—the small distinction between them is completely ‘masked’ by considerably larger ‘random’ distortion. Hence only signals that differ by not less than a certain definite value turn out to be discernible here. Since, in addition, the maximal level of our signals (defining the average signal power P) also cannot be unboundedly large, there can be only a finite number m of levels of signal values distinct from each other. A quantitative analysis of the situation arising here has been carried out by Shannon [21] (see also [24] or [25]), showing that in general the number m can be defined by the equation $m = \sqrt{1 + (P/W)}$. Thus, we arrive at the following expression for the capacity C of an arbitrary channel, transmitting continuously varying signals :

$$C = L_1 \log \left(1 + \frac{P}{W} \right), \quad L_1 = \frac{L}{2} \quad (*)$$

(where L_1 is some ‘universal’ characteristic of communication channels, irrespective of the signal to be transmitted).† The conclusion that stems from this remarkable formula is one of the most important contributions of information theory to general communication theory.

The deduced formula enables us to calculate easily the capacity of every concrete communication channel. In fact, apart from the engineering characteristics of a channel itself, it is also necessary to know the *signal to noise ratio* i.e., P/W . For teletransmission channels, C usually turns out to have an order of tens of millions of bits per second; for telephonic, phototelegraphic and radiotransmission channels, C varies from several thousands to several tens of thousands bits per second, and for telegraph channels C is of the order of tens or hundreds bits per second (see, for example, [115], [132] or [166]).

It is essential here that the existing channel capacity in all cases (except, perhaps, telegram) theoretically permit information transmission at a considerably higher rate than that achieved during ordinary practical transmissions. Thus, (say) in telegraphy, the information is transmitted usually at a rate not exceeding 75 bit/sec; in telephony, at a rate not exceeding 2,500 bit/sec; in television at a rate not exceeding 500,000 bit/sec. Hence all methods actually being employed

†Here we speak only of the capacity of a channel, transmitting *continuous* signals, since the case of the transmission of discrete signals in the presence of noise is especially examined in the next section.

at present for message transmission utilize as a rule just a small part of the available communication channel capacity. A higher capacity utilization prescribes the application of considerably more effective methods of encoding and decoding; this gives rise to many difficult problems, theoretical as well as purely applied, that are presently engaging the attention of a large number of workers all over the world (we shall speak of this in more detail in Sec. 4.5). Note that recent achievements in the field of the theory and practice of encoding and decoding now enable us in principle to enhance substantially the effectiveness of the use of communication channels: thus, in experimental transmissions especially organized by American scientists and engineers, an information transmission rate was successfully achieved that was of the order of 7,500–8,000 bit/sec over telephony (see, for example [25], p. 762; [94] or [199], p. 7) and 20,000,000 bit/sec over television (see [94]). However, even such information transmission rates nevertheless seem to be inadequate for future needs—the total amount of information to be transmitted through existing communication channels tends to increase every year in a majority of countries, and in the near future we may expect the evolution of new transmission models (say, video telephone), and also the emergence of two-way television communications between individual institutions in various cities and a massive use of direct digital data transmission to large centralized computer centres, which promise a significant acceleration of this process. Hence in present times in a number of world laboratories a start has been made to exploit fully new forms of communication channels having appreciably larger capacities, the foremost of them being the metallic and dielectric wave guide channels† with capacities of the order of $5 \cdot 10^8$ – $1 \cdot 10^9$ bit/sec and optical wave guides of glass fibres with a capacity of the order of 10^9 bit/sec per fibre. Such projects were discussed, in particular, at the International Conference on Communication Engineering held in Montreal in June 1971, at the International Information Theory Symposium at Tsakhkadzor, Armenian SSR, in September 1971 and at many other conferences related to communication engineering. Of course, the actual introduction of such new communication channels demands further circumventing of a large number of technical difficulties—but the very fact of the emergence of such studies is quite significant.

It is interesting to note that the mechanical concept of capacity can also be carried over completely to those 'communication channels' through which every living organism receives information from its sense organs. In fact, we have already described in Chapter 2 special psychological experiments, which show that the time required for assimilation of any information by the central nervous system is directly proportional to the amount of this information; thus the same

†The wave guides (radio and optical) are factually the pipelines through which the waves are propagated. The presence of an outer shell enables us to decrease strongly the noise level and together with this to use a very wide frequency band without creating interference with other communication channels.

laws are satisfied here as hold for all communication channels. Recently, some literature has also been published justifying the applicability of Shannon's formula (*) (p. 248) to the nervous communication channels in the human organism; however, at this point it is impossible to consider that the last word has been said on this question.

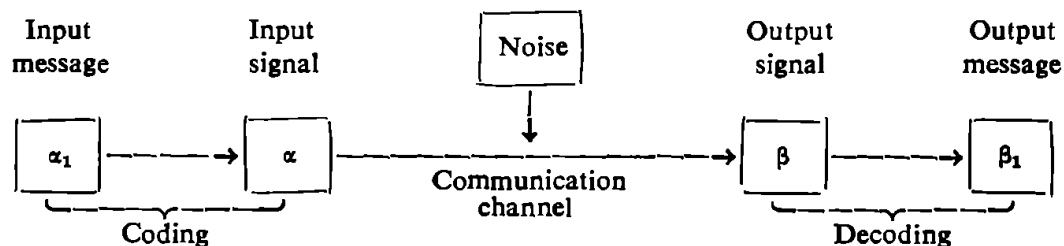
The capacity C of individual sense organs can be estimated quite roughly on the basis of the physiological data of their resolving power (i.e., the total number of objects to be distinguished by means of some sense organs) and the average time required for perception (i.e., the maximum frequency of the change of external influences during which these influences can all still be interpreted separately). This permits one to show, in particular, that the capacity of individual sense organs differs sharply; the human eye under favourable lighting conditions is obviously capable of receiving (and transmitting to the central nervous system) information at a rate of the order of millions (or tens of millions) of bits per second, whereas the ear receives information at a considerably slower rate—of the order of thousands of bits per second (see, for example, [109], [110], [114] and [156]). Such variances in the capacities can partially be explained by the sharp difference between the number of nerve fibres that serve, respectively, hearing and vision (according to modern physiological data, the number of 'aural nerve fibres' is of the order of 30,000, in contrast to roughly 800,000–900,000 'optic nerve fibres'). The sense of touch, in terms of its capacity to receive and transmit information, obviously lies somewhere between vision and hearing. However, it is worth noting that only a quite small part of the information transmitted to sense organisms can be assimilated consciously by the human brain; this clearly follows, for instance, from the data enunciated on p. 218 on the information reception rate in a conversation (it was observed there that when a conversation is rapid a part of the 'unsemantic' information is lost, since the listener has no time to reinterpret it). A detailed analysis of results concerning the maximum speed attainable in speech, reading, writing (shorthand) and so on, shows that in all cases a person is able to comprehend the information received only if the rate of its receipt does not exceed roughly 50 bit/sec (see, for example, [119] and [136]).† A quantity of the same order is also obtained in determining the amount of visual information to be assimilated by a spectator by a quick glance at changing figures projected on a screen [164]. Finally, especially designed experiments for the determination of the minimum physiological reaction time (see p. 56 et seq.), attainable under the most favourable conditions of reception, also show that the capacity of the human central nervous system is approximately equal to 30–40 bit/sec (see, [136] and [150]). Obviously, there still remains

†Recall also that in agreement with what is stated on p. 218 for a normal conversation just about half of the information to be received by the listener is contained in the written version of the speech; the rest of the information concerns the voice of the speaker, his emotions, 'insistence' stresses and so on.

much to be explored† with respect to further sharpening of these figures and clarifying their dependence on individual peculiarities of a person and his physical and mental condition. However, the fruitfulness of applications of the general ideas of information theory to the study of nerve activities of human beings and animals is no longer a suspect and is an established fact in its own right.

4.3.7. *A general scheme of information transmission through communication channels. Genetic information transmission*

In the present concluding subsection, we shall say a few more words on the general scheme of message transmission through communication channels, which formed factually our starting point in Section 4.1. The process of message transmission through an arbitrary communication channel can be schematically presented as follows :



In the case, (say) of the transmission of some texts through a telegraph channel, the input and output messages α_1 and β_1 are written in a definite (one and the same) language by means of the appropriate alphabet letters and can differ from one another only because of distortion in the transmission process, and the input and output signals α and β represent sequences of electric 'elementary' signals' (usually on and off currents). Thus, the coding and decoding operations consist here of the conversion of letter message α_1 into a sequence of 'elementary signals' α and in the reverse passage from the accepted sequence of elementary signals' β to the letter message β_1 . In a telephonic message transmission along a wire, α_1 has the character of sound, i.e., it is a sequence of air pressure fluctuations; the coding consists here of the transformation of these pressure fluctuations into electric current fluctuations, and decoding in the reverse transformation of accepted current fluctuations into sound. In the communication channels of modern electronic computers, the input signal α_1 is a definite sequence of numbers, the coding consists of its conversion into a definite sequence α of electric signals, directly fed into the computer, and decoding consists of the transformation of signals β received in the computer (representing the sum of 'input signals' α and the 'distortion in the input process') into an entirely new

†See, in particular, a survey of this question in [50] and references to the original literature listed there, which contain a multitude of data contradicting each other.

message β_1 , representing the solution of the problem we seek to solve with the aid of computer. Here β_1 in principle differs from α_1 and the conversion of α_1 into β_1 is the main goal of our communication channel. Similarly, in the case of the transmission of a visual 'image' through optical nerve fibres the 'messages' α_1 and β_1 differ sharply from each other—here α_1 consists of a collection of light waves of distinct wave lengths (i.e., distinct colours) and different amplitudes (i.e., intensiveness), and β_1 is a collection of stimulations from definite nerve cells (neurons) of the brain (the so-called 'visual neurons'), which are perceptible to us as a certain visual picture. The signal α in this particular case is a collection of electric pulses produced by the receptors of light (cones and rods of the retina) in the eye, and the coding consists of the conversion of light into such impulses, which so far have not been well studied. The decoding consists here of the transition from electric impulses β , reaching up to the brain through nerve fibres, to the stimulations of neurons β_1 , but its details are still considerably less known than those of the coding.

The general description of an arbitrary noisy communication channel and the determination of the theoretical limitations of the opportunities of using such a channel in information transmission, are examined in Section 4.4; and the concluding section (Section 4.5) forms an introduction to the extensive theory of optimal coding and decoding of discrete messages transmitted over noisy communication channels. However, we may only remark here that in many cases even the study of the 'alphabet' itself in which the messages α_1 and β_1 are written, and of the nature of 'elementary signals' α to be transmitted, is of great interest and not at all straightforward. The most striking example in this connection is the problem of transmitting *genetic* information, whose successful study is traced to a number of most outstanding scientific achievements during the last three decades.

In view of the general scientific importance of this topic and its intimate relation to the general formulation of the information transmission problem, it is appropriate to dwell here upon the related results in some detail. The 'communication channels' associated with the heredity phenomenon play a primary role in the very existence of organic life. Through these channels vast and extremely important information is constantly transmitted with striking precision. On Earth in all nearly 2 million individual species of animals and plants are recorded—and over the 'communication channels' under consideration signals are transmitted to indicate precisely what particular species ought to grow from a single embryonic cell. The information transmitted here is not at all restricted to just a single indication of species—it contains also sufficiently comprehensive data concerning the peculiarities of the structure of the species and, in addition, data concerning the hereditary singularities of an individual organism developed from a given cell. All this information is preserved somewhere in the extremely small volume of the nucleus of the embryonic cell and is transmitted through some sufficient complex pathway to the substance ('cytoplasm') of both the primary cell and all other cells that are produced in division processes originating at a given cell; it is preserved even in the process of subsequent reproduction of succeeding generations of similar species.

The construction of appropriate communication channels and the methods of information transmission over them seemed to be quite mysterious until recently. It was hardly possible to

foresee the rapid developments in this field that were linked to the spectacular advancements of molecular biology in the period following the last world war. A central role in this regard was played by the discovery of the fundamental importance of the enormous chain-like polymer molecules of the so-called *deoxyribonucleic acid* (abbreviated DNA), arranged in the chromosomes of the cell nucleus. It is known that these molecules consist of long alternating chains of carbohydrate and phosphate groups of identical composition. To each carbohydrate group there is also attached the side group from a collection of four standard bases called *adenine*, *guanine*, *cytosine* and *thymine*. All distinctions admissible in the DNA molecules are restricted to those concerning the successive interchange of corresponding bases (which, for brevity, may be denoted by their first letters *A*, *G*, *C* and *T*, or may also be just numbered by the digits 0, 1, 2 and 3). Thus, the original 'message' α_1 is preserved here in the chromosomes of the cell nucleus and written in a 'four-letter alphabet' of DNA molecules. One DNA molecule can store in a chromosome several tens of thousands or even more carbohydrate groups (and, consequently, also bases), and the number of individual chromosomes in a cell nucleus can amount to several tens; thus, the amount of information that can be stored in a chromosome is of the order of

$$\log 4^{100,000} = 200,000 \text{ bits}$$

(or even more). Thus, the amount of information that can be stored surpasses in abundance all transmitted data inherited by us.

In fact, the chromosome structure is still slightly more complex—usually chromosome includes not a single, but a *double* strand of DNA, composed of two such molecules, which are condensed in the form of two helixes coiled in the opposite directions around one (not actually existing) cylinder. These two DNA molecules are not identical but are 'complementary'—adenine in one of them always corresponds to thymine in the other and guanine to cytosine; the corresponding pair bases arranged on the cylinder surface opposite to each other are linked by a comparatively weak hydrogen bond. Such 'double helix' chromosome structure plays a key role in the process of their replication during cell division ('mitosis'), when each of the two new (daughter) cells reproduces for itself a set of chromosomes identical to that possessed by the original (parent) cell; this process is apparently related to the 'uncoiling' of the two DNA strands entering a chromosome, during which the two long DNA molecules get separated from each other and then each attaches itself to one more 'complementary' chain, forming an independent double helix. The resulting transmission of information from the parent cell to daughter cells plays a fundamental role in all biological phenomena; here the set of chromosomes (DNA chains) of the parent cell plays the role of the input 'message' α_1 to be transmitted and that of the two new daughter cells serves as the 'output message' β_1 . The 'output message' β_1 is obtained here directly from the 'input message' α_1 and this makes superfluous the problem of coding and decoding of 'messages'. At the same time the question of 'noise' in our communication channel is unusually important, because the distortions arising as a result of such 'noise' (caused, say, by radioactive irradiation of a cell) represent variations in hereditary characteristics ('mutation'), which occupy a central position in the process of the evolution of organic species.

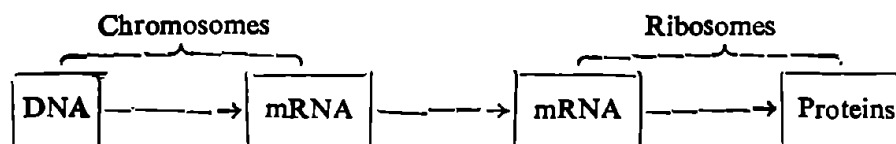
We now pass on to the information transmission from a chromosome to the 'body' (= 'cytoplasm') of the cell, which determines the development of any living creature, from one-celled organisms to man, from a single embryonic cell. An important role in all vital functions of an organism is assigned to the *proteins*, in particular to *enzymes*, which control all biochemical reactions that take place in living organisms. Protein synthesis takes place in the so-called *ribosomes*—small formations within the cytoplasm of a cell. The structure of protein molecules is also quite simple—all proteins are constructed from roughly 20 different *amino acids*, interchanging in a definite order along the linear protein molecules. These amino acids are listed in the accompanying table together with the abbreviations of their names that

are accepted in biochemistry. Each protein has its own characteristic sequence of amino acids ranging between 100 to 300 or more.

<i>Amino Acids</i>	<i>Abbreviations</i>	<i>Amino Acids</i>	<i>Abbreviations</i>
Alanine	ala	Leucine	leu
Arginine	arg	Lysine	lys
Asparagine	asn	Methionine	met
Aspartic acid	asp	Phenylalanine	phe
Cysteine	cys	Proline	pro
Glutamic acid	glu	Serine	ser
Glutamine	gln	Threonine	thr
Glycine	gly	Tryptophan	try
Histidine	his	Tyrosine	tyr
Isoleucine	ilu	Valine	val

Thus, it can be said that ribosomes serve as the receiving end ('output') of our communication channel; the 'output message' β_1 in this case is represented by proteins and it is written in a 'twenty-letter alphabet' of amino acids. It further remains to clarify just that how the transfer of information from DNA to proteins takes place, and in particular, what ought to be understood by the 'input signal' α and 'output signal' β .

A completely satisfactory answer can be given at present to the latter question. A key role in the process of information transmission from chromosome DNA to protein molecules is played by still another nucleic acid, the so-called ribonucleic acid (abbreviated RNA). The structure of RNA is closely similar to that of DNA, only the carbohydrate group is somewhat different here and thymine is replaced by a different base *uracil*, varying slightly from thymine in chemical composition. Thus, an RNA molecule can be considered as a 'signal' encoded with the aid of four 'elementary signals' *A, G, C* and *U* (or 0, 1, 2, and 3') that are quite similar to the 'letters' of the original 'message' *A, G, C* and *T*. Along the DNA molecules of chromosomes, as along a certain 'template', definite linear molecules of RNA (the so-called 'messenger' RNA or mRNA) are synthesized, which subsequently separate from the cell nucleus and penetrate into the ribosomes; these mRNA molecules play an important role in the process of protein synthesis. Thus, the general scheme depicted on p. 251 for information transmission through communication channels has the following form in the case considered:



Here the role of the 'input message' α_1 and 'output message' β_1 is assigned to the DNA and proteins, respectively, and that of the 'input signal' α and 'output signal' β to the molecules of mRNA,

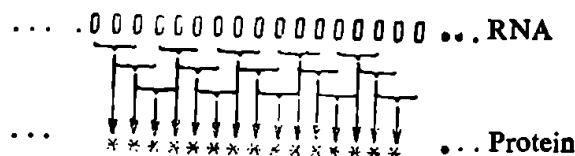
In accordance with the above scheme the 'transmitted message' α_1 is written in a 'four-letter alphabet' and the 'received message' β_1 in a 'twenty-letter alphabet' so that for our communication channel the number m of elementary signals entering its 'input' and the number r of elementary signals received at its 'output' are distinct ($m = 4$, and $r = 20$). Moreover, the 'codes', in which the 'signals' α and β are written, have four 'elementary signals.' As to the coding and decoding operations, i.e., the conversion of the 'message' α_1 into the 'signal' α and the 'signal' β into the message' β_1 , persuasive studies have been undertaken comparatively recently. Of the two operations enumerated above, 'coding' is naturally much more elementary (and hence also of less interest). In fact, coding involves the simple transformation of a sequence of four interchanging 'letters' A, G, C and T into a sequence of four 'elementary signals' A, G, C and U . Here it is possible to indicate many simple and easily realizable coding systems: thus, for instance, the singular 'complementarity' of specific base pairs manifested, in particular in the structure of 'double' DNA molecules, predicts a scheme, in which guanine 'produces' cytosine, cytosine—guanine, thymine—adenine and adenine produces, uracil. Apparently, such express coding is widely used in nature, though possibly it is not fully universal.†

Decoding is of considerably more interest in our case, since it consists of the intricate passage from the 'four-letter language' of mRNA to the 'twenty-letter language' of proteins. We have particularly this operation in view when we speak of a 'genetic code.' It is clear that one mRNA base, which can take in all *four* 'values' A, G, T and U , can by no means contain complete information on one of the *twenty* possible amino acids. Hence, it is necessary to consider that one amino acid is determined by a sequence of *several* adjoining bases in an RNA molecule: such a base sequence, 'coding' one letter of the amino acid alphabet, is usually called a *codon*. Since the number of distinct sequences of *two* RNA bases equals $4 \times 4 = 16$, i.e., it is less than the number of different amino acids, a codon must contain *not less than three* bases; *three* bases in a codon, however, suffice, since the number of all possible triplet bases equals $4 \times 4 \times 4 = 64$, which is far more than twenty amino acids.

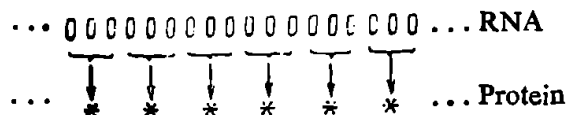
The first hypothesis on the nature of a genetic code was suggested in 1954 by the well-known American physicist and astrophysicist Gamow [100]. Gamow postulated that a given (let us say first) amino acid in a protein chain is determined by a certain *triplet* of successive RNA bases, (say) the first, second and third bases, and the following (second) amino acid by a triplet that is shifted by one, i.e., by the second, third and fourth bases; similarly the third amino acid is determined by a triplet that is shifted by two bases and so on. Such a code with partially overlapping codons is called an 'overlapping code' (see the scheme on the next page, where bases are denoted by small ovals and amino acids by asterisks). It was assumed here also that an amino acid protein depends only on the *composition* of the corresponding codon, but *not on the order* of the individual bases in the codon. The main argument that prompted Gamow to use this hypothesis was that the number of triplets distinct in composition that can be formed from four bases is given by

$$\begin{array}{ccccc} \binom{4}{3} & + & 2\binom{4}{2} & + & \binom{4}{1} & = 20 \\ \text{the number of triplets} & & \text{the number of triplets} & & \text{the number of triplets} & \\ \text{of mutually distinct} & & \text{containing two identi-} & & \text{of three identical bases} & \\ \text{bases} & & \text{cal bases} & & & \end{array}$$

†Thus, for instance, there are many viruses, in which the long RNA molecules replace DNA molecules as a primary genetic material. Hence the 'input message' α_1 is written here from the very start in the 'alphabet' A, G, T, U ,



The 'overlapping code' of Gamow was found to be not in agreement with reality, and the same was true of the 'nonoverlapping composition code' (see the accompanying scheme) sug-



gested by Gamow and Yčas [102] (in this code an amino acid protein is determined also by the *composition* of a corresponding codon, but codons do not overlap with each other). However, the clarity in Gamow's formulation of the main problem of protein synthesis in a living cell played a significant role in further development of this branch of molecular biology. The problem can be described as that of the 'translation' of a signal β written in the four-letter RNA language into a message β_1 written in a twenty-letter protein language, which is consistent with the experimental data.

At one time the idea of a 'code without commas' introduced by Crick and his group [88] competed with the 'composition code' of Gamow and Yčas. Such codes for a reasonably long time were discussed widely by a number of scientists (see, for example, the paper [104] written jointly by the mathematicians Golomb and Welch and the geneticist Delbrück). Here the term 'code without commas' is understood slightly different from that on p. 140—where this was used to mean an arbitrary uniquely decipherable code (each uniform code consisting of only three-letter codons evidently satisfies the last condition). But if we assume that a code is nonoverlapping, then it is not clear how the end of one codon and the start of next is discerned. In fact, the same sequence of bases, say ... *AGGCTCA* ..., can be divided variously into three-letter 'codons'; it can be 'read' either as ... (*AGG*) (*CTC*) (*A* ...), or as ... *AG*) (*GCT*) (*CA* ...), or as ... *A*) (*GGC*) (*TCA*) There are at least three different possible ways to avoid the uncertainties that arise. In principle, there can be some particular 'initiation mark' indicating the starting point of a codon sequence.† It is also possible that a special base sequence exists (it perhaps contains a larger or smaller number of bases than the codons do) that separates the individual codons from each other—such base sequence is then deciphered as a 'comma' separating the 'words' (codons) from each other. Finally, communication theory specialists are also aware of 'codes without commas' such that an arbitrary sequence of 'letters' (in our case, DNA bases) admits just *one* possibility for a meaningful reading, while any other way of dividing this letter sequence into individual 'words' leads to a sequence of meaningless letter combinations.

It is clear that a 'code without comma' thus defined must be 'incomplete'—there must exist in it letter sequences which designate no 'words' (constitute no codons). Accepting that every

†Let us note here that apparently this particular variant is realized in nature. There are special 'initiation' and 'termination' marks indicating the initiation and termination of a 'gene'—a base sequence that codes a specific protein produced in a given cell. In many cases, different genes are also divided by a definite series of bases which contain no information about any amino acid of the cell but play a distinct biological role,

codon consists of three bases (a *triplet* code), it is easy to determine the greatest possible number of intelligible codons. It is clear that a 'triplet' consisting of three identical 'letters' (bases), (say) *A A A*, cannot have a sense, because otherwise a long sequence of corresponding 'letters' ... *A A A A A A A A* ... can be read sensibly, starting from *any* place. The remaining $64 - 4 = 60$ distinct triplets can be divided into 20 groups of 3 triplets each, obtained from each other by the 'cyclic rearrangement of letters' (bases); examples of such triplets are *AGC*, *GCA* and *CAG*, or *CCT*, *CTC* and *TCC*. It is evident that out of these three triplets just one can make sense, because otherwise it would also be impossible to determine uniquely from what place it is necessary to start the reading of codons in a long sequence of identical triplets of one of these forms. Thus the largest possible number of sensible codons in the case of a triplet code without comma cannot exceed $60 \div 3 = 20$. It can also be shown that in fact it is exactly *equal* to 20. This fact provided Crick and the researchers sharing his viewpoint with a strong argument in favour of the hypothesis that genetic code is a 'code without comma.'

The solution of the problem of the structure of 'genetic code' was, however, obtained not on a writing table but directly in the laboratories. In the early sixties, a group of biochemists led by Marshall W. Nirenberg succeeded in showing that a synthesis of protein-like amino acid chains can be accomplished experimentally even in a cell-free system (i.e., in the absence of living cells). The system was made by breaking open cells of some bacillus, extracting from them ribosomes and all the basic components of a cytoplasm medium and adding the obtained material to a synthetic RNA of a definite composition, which in the process of protein synthesis enacts the role of a messenger RNA of the living cell. In the first such experiment carried out by Marshall Nirenberg and Heinrich Matthaei, the synthetic RNA contained only one repetitive *uracil* base; here was observed the synthesis of an artificial protein consisting of repetitive amino acid *phenylalanine* (phe). Thus the RNA ... *U U U U U U U U* ... generated the proteins ... *phe phe phe* ... , which implies that if a code is a triplet, then the amino acid *phe* must correspond to the codon *U U U*. Similarly, it was shown that the amino acid *proline* (pro) corresponds to the codon *C C C*.

During the sixties a vigorous 'attack' was launched on the problem of the genetic code by the numerous biochemical laboratories of the world. Among the participants in this campaign, besides Nirenberg and his associates (among whom Philip Leder played a very important role), we must mention the India-born scientist H. Gobind Khorana and the Mexican scientist Severo Ochoa, both working in the USA. We shall not dwell upon this at length here, and refer the interested reader to the relatively old surveys of Gamow, Rich and Ycas [101], describing the initial stages of the endeavour to decipher the genetic code, to the (also sufficiently early) popular articles of Crick and Nirenberg [87], intended for the non-specialists, and particularly to the self-contained monograph of Ycas [175], listing more than 800 references. The researches of many scientists have established that the genetic code is indeed a *triplet* and *nonoverlapping*;† that it is '*degenerate*' in the sense that *several* different codons directly correspond to a particular amino acid; that there are '*nonsense*' (i.e., carrying no genetic information) triplets, which

†The genetic code is non-overlapping in the sense that two successive codons of a gene do not overlap but occupy two adjacent base triplets. However, it was discovered recently (see [72]) that in the middle of a gene an 'initiation mark' can also appear which is shifted by one or two bases relative to a codon beginning of the primary gene. This initiation mark indicates the beginning of a new gene which overlaps with the first one; all the codons of the new gene are shifted in relation to codons of the first gene, i.e., each new codon overlaps two adjacent codons of the first gene. This discovery is quite interesting, but it does not affect any foregoing result related to the biological code.

in general are not codons in the sense that they do not correspond to any amino acids.†

The accompanying table from [175] shows the genetic code as interpreted by contemporary scientists (a dash in the left column implies that the related triplet is not a codon).

<i>Codon</i>	<i>Amino Acid</i>	<i>Codon</i>	<i>Amino Acid</i>	<i>Codon</i>	<i>Amino Acid</i>	<i>Codon</i>	<i>Amino Acid</i>
<i>UUU</i>	phe	<i>UCU</i>	ser	<i>UGU</i>	cys	<i>UAU</i>	tyr
<i>UUC</i>	phe	<i>UCC</i>	ser	<i>UGC</i>	cys	<i>UAC</i>	tyr
<i>UUA</i>	leu	<i>UCA</i>	ser	<i>UGA</i>	—	<i>UAA</i>	—
<i>UUG</i>	leu	<i>UCG</i>	ser	<i>UGG</i>	try	<i>UAG</i>	—
<i>CUU</i>	leu	<i>CCU</i>	pro	<i>CGU</i>	arg	<i>CAU</i>	his
<i>CUC</i>	leu	<i>CCC</i>	pro	<i>CGC</i>	arg	<i>CAC</i>	his
<i>CUA</i>	leu	<i>CCA</i>	pro	<i>CGA</i>	arg	<i>CAA</i>	gln
<i>CUG</i>	leu	<i>CCG</i>	pro	<i>CGG</i>	arg	<i>CAG</i>	gln
<i>AUU</i>	ilu	<i>ACU</i>	thr	<i>AGU</i>	ser	<i>AAU</i>	asn
<i>AUC</i>	ilu	<i>ACC</i>	thr	<i>AGC</i>	ser	<i>AAC</i>	asn
<i>AUA</i>	ilu	<i>ACA</i>	thr	<i>AGA</i>	arg	<i>AAA</i>	lys
<i>AUG</i>	met	<i>ACG</i>	thr	<i>AGG</i>	arg	<i>AAG</i>	lys
<i>GUU</i>	val	<i>GCU</i>	ala	<i>GGU</i>	gly	<i>GAU</i>	asp
<i>GUC</i>	val	<i>GCC</i>	ala	<i>GGC</i>	gly	<i>GAC</i>	asp
<i>GUA</i>	val	<i>GCA</i>	ala	<i>GGA</i>	gly	<i>GAA</i>	glu
<i>GUG</i>	val	<i>GCG</i>	ala	<i>GGG</i>	gly	<i>GAG</i>	glu

4.4. Transmission of information over noisy channels

In sections 1 and 2 of this chapter we sketched briefly via an example from telegraphy the basic concepts and results from the general theory of transmission of information over a communication channel. It was, however, always implied there that signals were transmitted over a communication channel *without any distortion*. But in practical communications this never occurs; there is always a possibility of some noise, which causes distortion of the signals in the trans-

†But, nevertheless, they are genetically important (in particular, they can indicate the beginning and the end of a gene, i.e., can play the role of initiation or termination marks). In this context, see [175, Chap. VIII].

mission process. A passing reference to this was already made in Section 4.3 in connection with the analysis of the performance of a communication channel transmitting continuous signals (see pp. 247–248). In this section we revert to the simple scheme of discrete communication channels considered in Sections 4.1 and 4.2, i.e., we assume that only a *finite number* of distinct ‘elementary signals’ of constant duration are transmitted over the channel. (The simplest case is of course that in which there are only two distinct signals, on-current and off-current.) However, contrary to Sections 4.1 and 4.2, we shall no longer ignore the influence of noise. This means that we shall take note of the possibility that the given elementary signals, in consequence of the distortion induced by noise, may be erroneously interpreted at the receiving end as some different elementary signal (for example, on-current may be misinterpreted as off-current). Let us now consider the application of information theory to this more complex (but, on the other hand, also more real) case of information transmission.

Following Section 4.2, we assume for simplicity that the successive ‘letters’ of the message are mutually independent, and the n letters of the alphabet are characterized by definite probabilities p_1, p_2, \dots, p_n of the appearance of some letter at any place. We consider a communication channel, in which m different elementary signals A_1, A_2, \dots, A_m are used for transmission, and L such signals are transmitted per unit of time (i.e., the duration of one signal is $\tau = 1/L$). Then, according to the main results of Section 4.2, *it is possible to transmit information across a noiseless communication channel at a rate arbitrarily close to the quantity*

$$v = \frac{C}{H} \text{ letters per unit time}$$

(where

$$C = L \log m$$

is the capacity of the communication channel, and

$$H = -p_1 \log p_1 - p_2 \log p_2 - \dots - p_n \log p_n$$

is the entropy of a single letter of the message to be transmitted); however, *a transmission rate exceeding v can never be attained* here. In order to attain a transmission rate extremely close to v , the only requirement is to partition the message into sufficiently long blocks and encode individual blocks by means of, for instance, the Huffman optimal code or some nearly optimal code (say, the Shannon-Fano code, or any code with word lengths l_i such that $-\log p_i / \log m \leq l_i < -\log p_i / \log m + 1$). In other words, the prescription here is to make use of a code for which the redundancy in the encoded message is the least possible or at least sufficiently close to it,

The foregoing description of a noisy communication channel can also be generalized by admitting that noise may sometimes so distort the transmitted signal that at the output it cannot be identified with any of the transmitted m elementary signals A_i . In order to take care of such a possibility it is expedient to assume that at the output there may be obtained not necessarily those very m elementary signals A_1, A_2, \dots, A_m which were transmitted through the channel, but some other r (where r can be greater than, less than, or equal to m) elementary signals B_1, B_2, \dots, B_r (all or some of which may be distinct from A_1, A_2, \dots, A_m , see Example 4° below). In this case, noise can be characterized by the mr positive numbers

[illegible]

Let us now assume that $p(A_1), p(A_2), \dots, p(A_m)$ are, respectively, the probabilities of the transmission of the signals A_1, A_2, \dots, A_m (here, obviously, $p(A_1) + p(A_2) + \dots + p(A_m) = 1$). In such case, the experiment β consisting of the determination of what specific signal is transmitted has the entropy $H(\beta)$ given by

$$H(\beta) = -p(A_1) \log p(A_1) - p(A_2) \log p(A_2) - \dots - p(A_m) \log p(A_m).$$

†Generally speaking, we can generalize slightly further even this parametrization by assuming that an arbitrary (i.e., (say) infinite or even continuous) set of signals B can be obtained at the output. We can carry over to this case almost all results indicated below; however, a number of equations now appear to be more complex. Due to this reason, the indicated generalization of the notion of a communication channel will not be drawn upon in the following.

that β has the outcome A_i (where $i = 1, 2, \dots, m; j = 1, 2, \dots, r$), is just equal to $p_{A_i}(B_j)$. The average amount of information about experiment β contained in the experiment α is

$$I(\alpha, \beta) = H(\beta) - H_{\alpha}(\beta),$$

where $H_{\alpha}(\beta)$ is the conditional entropy defined by the equations on pp. 61–63 (with the replacement of k and l by m and r in these equations). It is clear that the information $I(\alpha, \beta)$ never exceeds the entropy $H(\beta)$ of β , since $H(\beta)$ is equal to the maximum information about β that can be obtained from any experiment (this information is contained, for example, in β itself). The information $I(\alpha, \beta)$ equals the entropy $H(\beta)$ if and only if the outcome of β is uniquely defined by the outcome of α , i.e., if and only if the received signal allows us to determine uniquely what signal has been transmitted (from a practical viewpoint this means that here noise does not affect the reception). The information $I(\alpha, \beta)$ is zero when α does not depend on β (i.e., when the signal received does not at all depend on what signal is transmitted—in other words, when, because of quite strong noise, the transmission of information factually does not take place).

We now recall that the *channel capacity* C of a noiseless communication channel is defined in Section 4.2 as the *greatest amount of information that can be transmitted through this channel per unit of time* (see p. 173). Let us extend this definition to the case of noisy channel. For such a channel, the average amount of information conveyed by one elementary signal received at the channel output is given by

$$I(\alpha, \beta) = H(\beta) - H_{\alpha}(\beta),$$

i.e., it depends on the probabilities $p(A_1), p(A_2), \dots, p(A_m)$ that the signals A_1, A_2, \dots, A_m are transmitted. Let

$$c = \max I(\alpha, \beta)$$

be the *maximum* value of $I(\alpha, \beta)$ which can be attained by the variation of probabilities $p(A_1), p(A_2), \dots, p(A_m)$, and suppose that this value is achieved for the values $p^0(A_1), p^0(A_2), \dots, p^0(A_m)$ of these probabilities (examples of the explicit evaluation of the quantity c and the probabilities $p^0(A_1), p^0(A_2), \dots, p^0(A_m)$ will be given below). The quantity c is defined as the largest amount of information that can be obtained at the output when a single elementary signal is transmitted. If it is desired to obtain the greatest amount of information transmitted during a definite time interval (say, in a given unit of time), then it is natural to act as follows. During the indicated time interval, we always select the values of transmitted elementary signals with the same probabilities $p^0(A_1), p^0(A_2), \dots, p^0(A_m)$ regardless of what specific signals were transmitted previously. (For the justification of this method, see the text in small print on pp. 297–298,

where it is rigorously proved that for any transmitted sequence of mutually dependent signals the total amount of transmitted information cannot exceed the amount of information transmitted when the best independent signals are used.) In such a transmission each elementary signal transmits c units of information, i.e., the amount of information conveyed per unit time is given by

$$C = Lc = L \max I(\alpha, \beta).$$

This quantity C is called the *capacity of a noisy channel*. Since the maximum of $I(\alpha, \beta)$ cannot exceed $H(\beta)$ and $H(\beta)$ is never greater than $\log m$ (see pp. 48–49), it is clear that the capacity of a noisy channel is never greater than that of a noiseless channel through which the same number of elementary signals can be transmitted per unit time and which uses the same number of distinct signals. Consequently, noise can only decrease the channel capacity, which agrees well with the inference dictated by common sense.

Examples

1°. Let us begin with the case when $r = m$, the signals B_1, \dots, B_r coincide with A_1, \dots, A_m and $p_{A_i}(A_j) = 1$ for $j = i$, and hence $p_{A_i}(A_j) = 0$ for $j \neq i$. Here, we always receive the *same* signal as that transmitted (noise does not affect the transmission or is even totally absent) and hence

$$H_{\alpha}(\beta) = 0 \text{ and } c = \max I(\alpha, \beta) = \max H(\beta) = \log m.$$

This maximum value is achieved, as we know, when all the possible signals to be transmitted are equally likely, so that here $p^0(A_1) = p^0(A_2) = \dots = p^0(A_m) = 1/m$. Thus in this case, $C = L \log m$. Hence it is seen that the definition of the capacity of a noiseless channel derived in Section 4.2 is a particular case of the more general definition considered here.

2°. Suppose that two elementary signals (say, the on-current A_1 and off-current A_2) can be transmitted through a communication channel and the same two signals A_1 and A_2 are obtained at the receiving end. Further, suppose that p and $1 - p$ are the respective probabilities of receiving any of the signals with and without error. We have

$$p_{A_1}(A_1) = p_{A_2}(A_2) = 1 - p, \quad p_{A_1}(A_2) = p_{A_2}(A_1) = p,$$

so that the table of conditional probabilities given on p. 260 here has the form

$$\begin{array}{cc} 1 - p, & p; \\ p, & 1 - p. \end{array}$$

The corresponding communication channel is called a *binary symmetric channel*;

it is schematically shown in Fig. 18, where arrow-heads indicate the received signals into which A_1 and A_2 may be transformed, and along the lines of the arrow are written the probabilities of the corresponding reception.

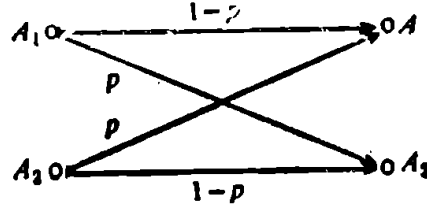


Fig. 18.

To evaluate the quantity c , we use the equality

$$I(\alpha, \beta) = H(\alpha) - H_{\beta}(\alpha).$$

From the table of conditional probabilities given above, it is seen that if A_1 is transmitted, then we obtain at the receiving end the same signal A_1 with probability $1 - p$ and the signal A_2 with probability p ; if, however, A_2 is transmitted, then we receive A_1 with probability p and A_2 with probability $1 - p$. Hence

$$H_{A_1}(\alpha) = H_{A_2}(\alpha) = -(1 - p) \log (1 - p) - p \log p = h(p),$$

where $h(p)$ is the function introduced on p. 49, and

$$H_{\beta}(\alpha) = p(A_1)H_{A_1}(\alpha) + p(A_2)H_{A_2}(\alpha) = h(p),$$

independently of the values of probabilities $p(A_1)$ and $p(A_2)$ [because we always have $p(A_1) + p(A_2) = 1$]. Consequently, in this case $H_{\beta}(\alpha)$ does not at all depend on $p(A_1)$ and $p(A_2)$ and for calculating

$$c = \max I(\alpha, \beta) = \max [H(\alpha) - H_{\beta}(\alpha)]$$

it is required only to determine the maximum value of $H(\alpha)$. But $H(\alpha)$ is the entropy of an experiment α with two possible outcomes, which can in no way exceed 1 bit (see p. 49). On the other hand, the value $H(\alpha) = 1$ is certain to be attained when $p(A_1) = \frac{1}{2}$, $p(A_2) = \frac{1}{2}$, since in that case clearly both outcomes of α have identical probabilities [in the general case, these probabilities obviously equal

$$q(A_1) = p(A_1)(1 - p) + p(A_2)p,$$

and

$$q(A_2) = p(A_1)p + p(A_2)(1 - p)].$$

Hence, in our case it follows that

$$p^0(A_1) = p^0(A_2) = \frac{1}{2},$$

$$c = 1 + (1 - p) \log (1 - p) + p \log p = 1 - h(p),$$

and

$$C = Lc = L[1 - h(p)].$$

We have thus derived explicit equation that determines the dependence of the capacity of a binary symmetric channel on the probability p of an erroneous transmission. The graph of the function $C(p)$ is given in Fig. 19. This function attains a maximum ($= L$) when $p = 0$ (i.e., in the absence of noise) and when $p = 1$ (i.e., in the case of noise which transforms *each* transmitted signal A_1

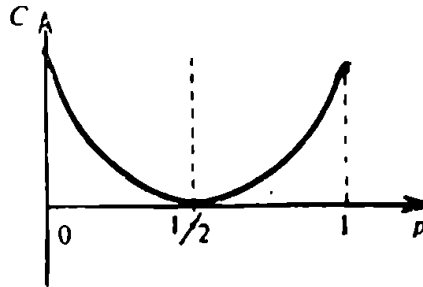


Fig. 19.

into A_2 and vice versa; it is obvious that such noise does not hinder us from understanding what signal has been transmitted). It is also clear that, generally, if $p > \frac{1}{2}$ then we can always replace in the received message every signal A_1 by A_2 and vice versa; this transforms the given channel into a communication channel with error probability $1 - p < \frac{1}{2}$. Hence it is obvious that the replacement of p by $1 - p$ cannot affect the value of the channel capacity C (this is seen also from the formulae obtained above), i.e., the graph of the function $C(p)$ must be symmetric in the line $p = \frac{1}{2}$. When $p = \frac{1}{2}$, the channel capacity C is zero; this is related to the fact that for $p = \frac{1}{2}$, *regardless of what signal is transmitted*, we get at the receiving end both signals A_1 and A_2 with probability $\frac{1}{2}$, so that the received signal contains no information about what signal has been transmitted.† When the values of p range between 0 and $\frac{1}{2}$ (or between $\frac{1}{2}$ and 1), we have a positive channel capacity less than L ; moreover, this channel capacity rapidly decreases for increasing p (when $p < \frac{1}{2}$) or $1 - p$ (when $p > \frac{1}{2}$). Thus, for instance, if $L = 100$, then for $p = 0.01$ (i.e., when out of 100 transmitted binary signals on the average one is received in error) $C \approx 92$ bits; when $p = 0.1$ (i.e., if 10 out of 100 signals suffer from noise distortion) $C \approx 53$ bits, and when $p = 0.25$ (i.e., if one fourth of all received signals are wrong) $C \approx 19$ bits.

3°. Let us now consider a more general example of communication channels, using m distinct elementary signals A_1, A_2, \dots, A_m , where the same signals are also obtained at the receiving end of the channel (i.e., $r = m$, $B_i = A_i$ for all i) and the probability of error-free transmission of each of these signals is $1 - p$, but in the case of erroneous transmission the transmitted signal may with identical probability, i.e., $p/(m - 1)$, be taken as any of $m - 1$ signals different

†In place of a communication channel, we can use equally successfully the result of the throw of a coin and consider the signal A_1 (resp. A_2) to have been received when the 'head' (resp. 'tail') turns up.

from it. The table of conditional probability assumes here the form

$$\begin{aligned} &1 - p, \frac{p}{m-1}, \frac{p}{m-1}, \dots, \frac{p}{m-1}; \\ &\frac{p}{m-1}, 1 - p, \frac{p}{m-1}, \dots, \frac{p}{m-1}; \\ &\frac{p}{m-1}, \frac{p}{m-1}, \frac{p}{m-1}, \dots, 1 - p; \end{aligned}$$

and the corresponding communication channel is called an *m*-ary symmetric channel. Let us again make use of the representation of $I(\alpha, \beta)$ in the form $H(\alpha) - H_\beta(\alpha)$. Then, obviously,

$$\begin{aligned} H_{A_1}(\alpha) &= H_{A_2}(\alpha) = \dots = H_{A_m}(\alpha) \\ &= -(1-p) \log(1-p) - (m-1) \times \frac{p}{m-1} \log \frac{p}{m-1}, \end{aligned}$$

and, consequently,

$$H_\beta(\alpha) = -(1-p) \log(1-p) - p \log \frac{p}{m-1}.$$

Thus, as in example 2°, it is again ascertained that $H_\beta(\alpha)$ is independent of the probabilities $p(A_1), p(A_2), \dots, p(A_m)$ and for finding the channel capacity the only requirement is to determine the maximum value of $H(\alpha)$. In complete analogy with Example 2°, this maximum value is found to be $\log m$ and is attained when all outcomes of experiment α (i.e., all possible values of received signals) are equally probable. The last condition is obviously satisfied when the probabilities $p(A_1), p(A_2), \dots, p(A_m)$ of sending the signals A_1, A_2, \dots, A_m also are all equal to each other. Hence

$$p^0(A_1) = p^0(A_2) = \dots = p^0(A_m) = \frac{1}{m},$$

$$c = \max I(\alpha, \beta) = \log m + p \log \frac{p}{m-1} + (1-p) \log(1-p),$$

and

$$C = L \left[\log m + p \log \frac{p}{m-1} + (1-p) \log(1-p) \right].$$

The graph of the function $C(p)$ (for $m = 4$) is given in Fig. 20. This function attains its maximum value $L \log m$ when $p = 0$ (in the absence of noise), and when p increases from 0 to $p = (m-1)/m$ it reduces smoothly to zero.

The fact that the channel capacity for $p = (m - 1)/m$ is zero is quite natural: in this case for any signal to be sent we can obtain each signal A_1, A_2, \dots, A_m with equal probability $1/m$ at the receiving end, so that no information is conveyed here regarding the signal to be sent. With further increase of p we again obtain (truly, not large) a positive channel capacity; in this case, if we receive

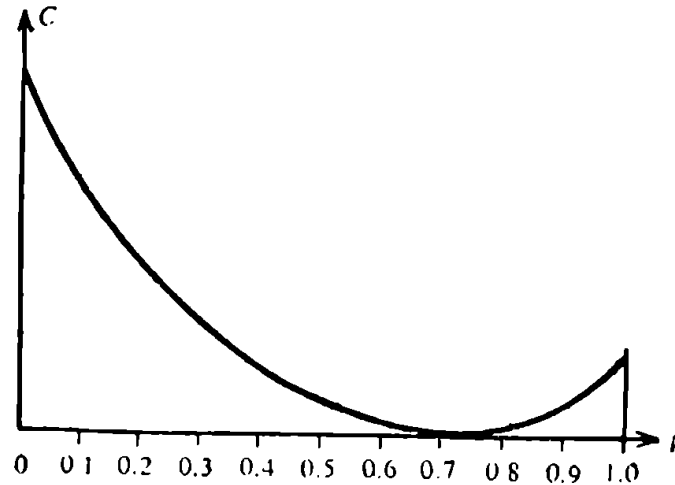


Fig. 20.

signal A_i , we can infer that the probability of the transmission of any signal *other than* A_i is larger than the probability of the transmission of A_i , i.e., we have nevertheless some information as to what specific signal is transmitted. Due to this fact, the channel capacity again increases when p increases from $(m - 1)/m$ to unity, namely it becomes $L \log [m/(m - 1)]$ for $p = 1$.

4°. Consider again a binary communication channel through which two signals A_1 and A_2 can be transmitted; however, it is now assumed that the signal obtained at the output may sometimes be interpreted as one of these two signals but occasionally the signal may be so distorted that it is completely impossible to identify it. In the latter case, it is appropriate to suppose that an altogether new signal A_3 is received, whose appearance can be interpreted as an event: the transmitted signal has been completely erased and cannot be deciphered (hence such communication channel is called the *binary erasure channel*). We shall confine ourselves here to the simplest case of a *binary symmetric erasure channel*, for which the probability of any of the transmitted signals A_1 and A_2 to be erased is one and the same number q (i.e., $p_{A_1}(A_3) = p_{A_2}(A_3) = q$); moreover, if the erasure does not arise, then both the signals A_1 and A_2 will be deciphered correctly at the output with one and the same probability $1 - p - q$, and with probability p they will be confused (i.e., signal A_1 will be confused with signal A_2 and vice versa). Thus, in the case of a binary symmetric erasure channel we have $m = 2$, $r = 3$ and the corresponding table of conditional probabilities

$p_{A_i}(B_j) = p_{A_i}(A_j)$ takes the form

$$\begin{array}{ccc} 1 - p - q, & p, & q; \\ p, & 1 - p - q, & q \end{array}$$

(see Fig. 21).

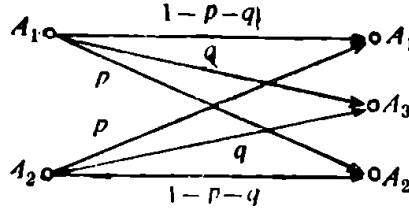


Fig. 21.

It is clear that no matter what signal is transmitted, we obtain at the receiving end the signals A_3 with probability q , whereas of the two remaining signals one has probability $1 - p - q$ and the other probability p . Consequently,

$$H_{A_1}(\alpha) = H_{A_2}(\alpha) = -(1 - p - q) \log (1 - p - q) - p \log p - q \log q,$$

and hence

$$H_{\beta}(\alpha) = -(1 - p - q) \log (1 - p - q) - p \log p - q \log q,$$

so that

$$I(\alpha, \beta) = H(\alpha) + (1 - p - q) \log (1 - p - q) + p \log p + q \log q.$$

Since experiment α can have three outcomes A_1 , A_2 and A_3 in the considered case, we have $H(\alpha) \leq \log 3$; hence

$$c = \max I(\alpha, \beta) \leq \log 3 + (1 - p - q) \log (1 - p - q) + p \log p + q \log q.$$

But *can* the entropy of α be equal to $\log 3$? It is easy to see that, in general, it *cannot*, whatever the probabilities $p(A_1)$ and $p(A_2)$ of signals A_1 and A_2 . In fact, the equality $H(\alpha) = \log 3$ is satisfied if and only if all the three outcomes of α are equally probable (i.e., all have probability $\frac{1}{3}$). In our case, however, the probability of outcome A_3 ('erasure') with any choice of $p(A_1)$ and $p(A_2)$ is equal to the given number q , which is the channel characteristic and can, of course, be quite different from $\frac{1}{3}$. Hence, the entropy of α has the form

$$H(\alpha) = -q_1 \log q_1 - q_2 \log q_2 - q \log q,$$

where q is fixed, but $q_1 = p(A_1)(1 - p - q) + p(A_2)p$ and $q_2 = p(A_1)p + p(A_2)(1 - p - q)$ are the probabilities of obtaining the signals A_1 and A_2 , respectively at the receiving end, which depend on the values of $p(A_1)$ and $p(A_2) =$

$1 - p(A_1)$. It is clear that $q_1 + q_2 = 1 - q$ for all values of $p(A_1)$ and $p(A_2)$. But it is easy to see that the maximal value of $-q_1 \log q_1 - q_2 \log q_2$, where $q_1 + q_2 = 1 - q$ (and q is a fixed number satisfying the obvious condition $0 < q < 1$), is attained when $q_1 = q_2 = (1 - q)/2$.† In addition, it is also easily verifiable that the values $q_1 = q_2 = (1 - q)/2$ are in fact possible: for this it is only necessary to suppose that $p(A_1) = p(A_2) = \frac{1}{2}$. This yields

$$p^0(A_1) = p^0(A_2) = \frac{1}{2},$$

$$c = \max I(\alpha, \beta)$$

$$= -(1 - q) \log \frac{1 - q}{2} + (1 - p - q) \log (1 - p - q) + p \log p$$

$$= (1 - q) [1 - \log (1 - q)] + (1 - p - q) \log (1 - p - q) - p \log p$$

and, hence,

$$C = L\{(1 - q) [1 - \log (1 - q)] + (1 - p - q) \log (1 - p - q) + p \log p\}.$$

The channel capacity C obtained depends on two numbers p and q , which characterize probabilities of different types of errors in the given case. It is easy

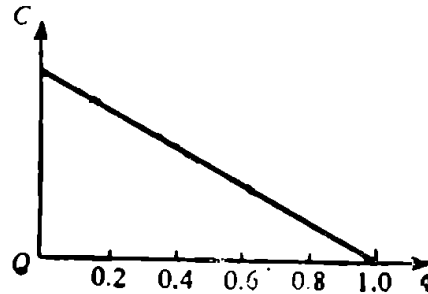


Fig. 22.

to show that C decreases for both increasing p and q (subject to the natural assumption that $p < \frac{1}{2}$). We further note that in an actual erasure binary com-

*In fact, by adding the constant term $(1 - q) \log (1 - q) = (q_1 + q_2) \log (1 - q)$ to $-q_1 \log q_1 - q_2 \log q_2$ and then multiplying the sum obtained by the constant factor $1/(1 - q)$, we get the expression

$$- \frac{q_1}{1 - q} \log \frac{q_1}{1 - q} - \frac{q_2}{1 - q} \log \frac{q_2}{1 - q},$$

which represents the entropy of an experiment with two outcomes having the probabilities $q_1/(1 - q)$ and $q_2/(1 - q)$. This entropy obviously takes its largest value when $q_1 = q_2$; consequently, the largest value of the original expression $-q_1 \log q_1 - q_2 \log q_2$ is also attained when $q_1 = q_2$.

munication channel the inequality $p < q$ is usually valid, i.e., the probability of the transmitted signal being so distorted that it becomes impossible to identify is usually larger than the probability of that distortion due to which it is found to resemble in form the second of the used signals. In a series of cases, the probability p is generally found to be so small that it can be completely ignored, i.e., it may be considered that the only possible detrimental distortion of signal due to noise is the one which makes the signal impossible to decipher at the output (i.e., at the output it gets 'erased'). If we agree to consider that $p = 0$, then the formula for the channel capacity C assumes the singularly simple form

$$C = L(1 - q)$$

(see Fig. 22). The preceding result is quite natural: when $p = 0$, out of L binary signals transmitted over our communication channel per unit time, on the average Lq signals are 'erased', i.e., do not convey any information, whereas the remaining $L(1 - q)$ signals are accurately deciphered at the receiving end, so that each of them transmits exactly 1 bit of information.

The circumstance that, in all the preceding examples the channel capacity C was attained when the probabilities of the transmission of any of the employed signals was taken to be equal to each other, is obviously accidental. This is explained merely by the fact that for simplicity of calculation in all these exam-

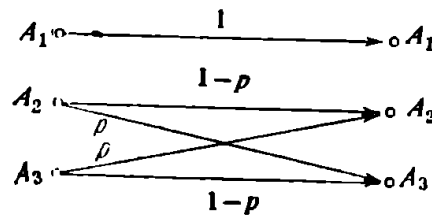


Fig. 23.

ples the table of channel conditional probabilities $p_{A_i}(B_j)$ was chosen to be quite symmetric. In order to illustrate that a different situation may also hold, we give one more result related to the following slightly more complex example first suggested by Shannon [21].

5°. Suppose that three elementary signals A_1 , A_2 and A_3 can be transmitted over a communication channel, where the first signal differs considerably from the other two and can always be unmistakably interpreted at the output of the channel, but each of the other two signals has probability $1 - p$ of being interpreted correctly and probability p of being misinterpreted as being its opposite. In other words, we consider that $m = r = 3$ and that the table of conditional probabilities $p_{A_i}(A_j)$ has the form

1,	0,	0;
0,	$1 - p$,	p ;
0,	p ,	$1 - p$

(see Fig. 23). Consequently,

$$H_{A_1}(\alpha) = 0,$$

$$H_{A_2}(\alpha) = H_{A_3}(\alpha) = -(1-p) \log (1-p) - p \log p = h(p),$$

and

$$H_{\beta}(\alpha) = [p(A_2) + p(A_3)] h(p),$$

$$I(\alpha, \beta) = -q(A_1) \log q(A_1) - q(A_2) \log q(A_2) - q(A_3) \log q(A_3) \\ - [p(A_2) + p(A_3)] h(p),$$

where $q(A_1) = p(A_1)$, $q(A_2) = p(A_2)(1-p) + p(A_3)p$ and $q(A_3) = p(A_2)p + p(A_3)(1-p)$ are the probabilities of the outcomes A_1 , A_2 and A_3 of α .

Note that $H_{\beta}(\alpha)$ does not depend on all the three probabilities $p(A_1)$, $p(A_2)$ and $p(A_3)$ but only on $p(A_2) + p(A_3) = 1 - p(A_1)$. Applying the arguments given in the footnote on p. 269, it is easy to show that with $p(A_1) = q(A_1)$ fixed the entropy $H(\alpha)$ (and, hence, also the information $I(\alpha, \beta)$) is the largest if the probabilities $q(A_2)$ and $q(A_3)$ (and, consequently, also $p(A_2)$ and $p(A_3)$) are equal to each other:

$$p(A_2) = p(A_3) = q(A_2) = q(A_3) = \frac{1 - p(A_1)}{2}.$$

The only requirement now is to determine for what value of $p(A_1)$ the expression

$$I(\alpha, \beta) = -p(A_1) \log p(A_1) - [1 - p(A_1)] \left[\log \frac{1 - p(A_1)}{2} + h(p) \right]$$

will be the largest, where p is a given nonnegative number not exceeding unity. This problem is quite complicated if only methods of elementary mathematics are used but is easy to solve with the aid of differential calculus.[†] It is found that the desired value of $p(A_1)$ is

$$p^0(A_1) = \frac{1}{1 + 2p^p(1-p)^{1-p}}.$$

Thus,

$$p^0(A_1) = \frac{1}{1 + 2p^p(1-p)^{1-p}},$$

[†]It is known that the point x of the segment $0 < x < 1$, at which the function

$$y = -x \log x - (1-x) [\log \{(1-x)/2\} - \log a]$$

(where $a = p^p(1-p)^{1-p}$ and all logarithms are taken to the base 2) takes its greatest value, coincides with that point at which the derivative of this function vanishes,

$$p^0(A_2) = p^0(A_3) = \frac{p^p(1-p)^{1-p}}{1 + 2p^p(1-p)^{1-p}}.$$

Substituting these probabilities in the expression for $I(\alpha, \beta)$ and multiplying the result by L , the number of signals transmitted per unit time, it is easy to find that the capacity of our communication channel is given by

$$C = L \log [1 + 2p^p(1-p)^{1-p}].$$

The graph of the function $C = C(p)$ is given in Fig. 24. For $p = 0$ this function takes its greatest value: it is easy to show that $p^p(1-p)^{1-p} \rightarrow 1$ as $p \rightarrow 0$ and, consequently, here $p^0(A_1) = p^0(A_2) = p^0(A_3) = \frac{1}{3}$ and $C = L \log 3$. This result is, indeed, obvious: for $p = 0$ we have a simple noiseless channel using three different elementary signals (see Example 1°). When p increases from 0 to $\frac{1}{2}$ the channel capacity C decreases, since in the transmission of the second or third signal a part of the information is lost because of the presence of noise.

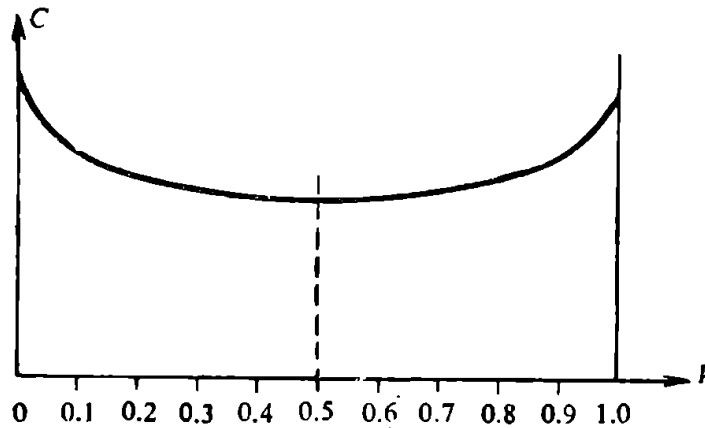


Fig. 24.

The probability $p^0(A_1)$ is, therefore, found to be slightly larger than $\frac{1}{3}$ (i.e., it is advantageous here to transmit the first signal more frequently than the second or third). For $p = \frac{1}{2}$ the channel capacity takes its smallest value, namely $C = L$ (since $(\frac{1}{2})^{\frac{1}{2}} \times (\frac{1}{2})^{\frac{1}{2}} = \frac{1}{2}$). For attaining this channel capacity, the first signal ought to be transmitted in half of the total cases ($p^0(A_1) = \frac{1}{2}$), and the second and third in the remaining half of the cases. Factually, the signals A_2 and A_3 ought to be considered here as one common signal, since at the output it is impossible to identify precisely one from the other and we simply say that either of them but not A_1 was transmitted. (Hence the case $p = \frac{1}{2}$ is equivalent to the case of a noiseless channel using two different signals.) When p further increases from $\frac{1}{2}$ to 1 the value of $C(p)$ again increases, where $C(p) = C(1-p)$ (by the same criterion as in Example 2°).

Another example of a communication channel for which the probabilities $p^0(A_i)$ are not equal among themselves can be obtained by assuming that $m = r = 2$

but that the probabilities of error in the transmission of the two given signals are not equal to each other (the case of a *binary asymmetric channel*). In this case, however, all formulas are found to be considerably more involved than those in the foregoing examples. We shall not, therefore, dwell upon them.

Let us now assume that the channel capacity C is known. In the case of a noiseless channel, as seen in Section 4.2, the value of C yields an accurate estimate of the greatest possible transmission rate of a message over a given channel: thus, whatever the coding method, this rate cannot exceed the quantity

$$v = \frac{C}{H} \text{ letters per unit time}$$

(where H is the entropy of a single letter of the message to be transmitted); a transmission rate arbitrarily close to v can, however, always be achieved. In a noisy channel, besides the transmission rate, we should also take into account the transmission accuracy characterized, for example, by the probability of error in determining every individual transmitted letter. It is easy to comprehend that *if the transmission rate v_1 (in letters per unit time) exceeds the quantity $v = C/H$ (where C as defined above is the channel capacity of a noisy communication channel!), then accurate transmission (permitting free of error reconstruction of all letters of the transmitted message) cannot take place by any means.* (This statement is, in fact, a loose formulation of the so-called *converse to the noisy coding theorem*, of which we shall speak more elaborately on p. 282 et seq.). Indeed, in an error-free transmission at a rate v_1 , the amount of information about the letters of a message transmitted through a channel per unit time is equal to the total amount of uncertainty of a v_1 -letter 'block', i.e., to the product $v_1 H$ (recall that the individual letters are assumed to be independent). Consequently, the amount of information transmitted per unit time about the code words at the channel input (i.e., about the outcomes of experiment β) cannot be all the more less than $v_1 H$ (see p. 89). But since $v_1 H > C$ when $v_1 > v = C/H$, it follows from the very definition of the quantity C that an error-free transmission of a message at a rate $v_1 > v$ letters per unit time cannot be accomplished. Starting from these reasonings, we can also evaluate precisely the lower bound of error probability attainable in the 'best possible' transmission of a message at a given rate $v_1 > v$ (cf. p. 282 et seq.)

It may be remarked further that if no restriction is imposed in general on the transmission rate of message, then in a majority of cases the probability of error in the determination of each transmitted letter can be easily *made as small as desired*. For this it usually suffices that every transmitted signal (or group of such signals) be repeated many times. It could, however, be expected that in order to obtain a quite low error probability it would be necessary that the transmission rate be substantially lowered (such a sharp fall in transmission rate will occur, in particular, if the probability of error is decreased by means

of the multiple repetition of all the signals). Strictly speaking, it may at first sight seem natural to think that every decrease in the error probability in the determination of the transmitted letters must inevitably be related also to a decrease in the transmission rate and that an indefinite decrease in error probability can by no means be attained without lowering indefinitely the transmission rate. It is, however, found that in reality this is not the situation. It has indeed been demonstrated by Shannon that *for every noisy communication channel we can always choose a particular code allowing us to transmit a message at a given rate arbitrarily close to*

$$v = \frac{C}{H} \text{ letters per unit time}$$

(but nevertheless necessarily slightly less than this quantity!) *and such that the probability of erroneous decoding of each transmitted letter is found to be less than any preassigned number ϵ* (say, less than 0.001, or 0.0001, or 0.000001). Obviously, the code of which we speak here will depend on ϵ and, as a rule, the smaller is the ϵ the more complex the code will be. The assertion set above in italics generalizes the fundamental coding theorem formulated in Section 4.2 and may be called the *fundamental noisy coding theorem*. A vital role in the proof of this theorem is played by the direct use of quite lengthy 'block' codes of a large number of letters; hence, the transmission of a message at a rate close to v and with a quite small error probability is associated with considerable delay in deciphering each transmitted letter.

Before we proceed further, it may be remarked here that, exactly as in the case of the fundamental noiseless coding theorem considered in Section 4.2, the restriction that individual letters of the text be mutually *independent* is in fact not essential. In what follows, we shall almost dispense with such a requirement and use only a particular related result, according to which, if N is sufficiently large, then out of n^N different N -letter blocks (where each letter can take n different values) only 2^{H^N} are 'probable' (and have nearly the same probability). For the case in which text letters are mutually dependent, this position does not hold. However, as remarked on p. 171, in this case also, subject to certain quite general conditions, among all possible N -letter blocks, where N is sufficiently large, it may be possible to extract a comparatively small portion of nearly equally probable N -letter blocks with a probability sum quite close to unity. The total number of 'probable' blocks of N mutually dependent letters is accordingly stated on p. 171 to be of the order of $2^{H_\infty N} \approx 2^{H^{(N)}}$, where $H^{(N)}$ is the entropy of an N -letter block and $H_\infty = \lim_{N \rightarrow \infty} H^{(N)}/N$ is the specific entropy of a single text letter.

Thus, if the text letters are dependent, then in general we need only replace throughout in the sequel the entropy H of one letter by the specific entropy H_∞ smaller than H . Moreover, in the case of a transmission rate v_1 exceeding $v = C/H_\infty$ letters per unit time, we can make use of the fact that the total amount of

information contained in $v_1 T$ letters of the transmitted text (where T is the transmission time), no matter what T , cannot be less than $v_1 T H_\infty$ bits. This implies, as can be easily shown, that the italicized statement on p. 274 remains valid even in the case of the transmission of a message whose letters are mutually dependent but the speed $v = C/H$ letters per unit time is replaced here by $v = C/H_\infty$ letters per unit time.

We shall now assume again for simplicity that the individual letters of the transmitted message are mutually *independent* (i.e., we shall everywhere use the customary entropy H of one letter and not the specific entropy H_∞). Unfortunately, even in this case a rigorous proof of Shannon's fundamental noisy coding theorem is quite tedious. Such a proof is absent in [21] which forms the starting point for all of information theory. In fact, Shannon [21] confines himself only to the exposition of some general arguments and a highly descriptive enunciation of the reasons for which this theorem must hold. A rigorous mathematical proof of this theorem was given later by Feinstein (see, for example [9]), whose underlying idea differs from the original reasoning of Shannon. A rigorous proof of this theorem, closely following the deductions briefly touched upon in [21] is contained in Shannon's [186], in which it is also shown that via the same path we can obtain stronger results, which we propose to take up below. In the present text, we start with certain very simple reasonings due to Shannon in order to initiate the reader into the fundamental coding theorem. Later, on p. 290 et seq., we also describe the rigorous methods of its proof, resting on deeper reasonings in [21]. In addition, taking into consideration its immense importance, we supplement (in small type) at the end of this section (see pp. 298-303) one more mathematical proof of this theorem for the particular case of a binary symmetric channel that is based on the idea similar to that followed by Feinstein.

Suppose that β is an experiment consisting of the choice (and subsequent transmission through a communication channel) of one of m elementary signals A_1, A_2, \dots, A_m with probabilities $p^0(A_1), p^0(A_2), \dots, p^0(A_m)$ that correspond to the maximum amount of information $I(\alpha, \beta)$ (i.e., for which the capacity of our channel is realized). Shannon's theorem says that there exists a method of encoding a message that enables us to carry out the transmission at a rate arbitrarily close to (but slightly less than !)

$$v = L \frac{c}{H} \text{ letters/unit time,}$$

where

$$c = H(\beta) - H_\alpha(\beta) = H(\alpha) - H_\beta(\alpha),$$

so that the probability of erroneously decoding the received message is small (smaller than an arbitrary preassigned small number). Since L elementary signals can be transmitted per unit time, the requirement for attaining such a trans-

mission rate is that the code word of N -letter 'block' contain nearly (but a few more than) $(H/c)N$ elementary signals. In fact, here the LT elementary signals transmitted during the large time interval T contain roughly $LT/(H/c)N = vT/N$ code words corresponding to a message of approximately vT letters.

It is known (see pp. 162–170) that in fact it need be of no concern to us that the code words of *all* $n^N = 2^{\log n \times N}$ distinct N -letter messages (where n is the number of alphabet letters) have a length close to $(H/c)N$ signals. Indeed, only 2^{HN} of these N -letter messages are 'probable'; as regards the remaining $2^{\log n \times N} - 2^{HN}$ messages, the total probability of their appearance for large N is quite small, and hence even if their code words were appreciably longer, this does not noticeably lower the transmission rate (remaining close to $L(c/H)$ letters per unit time). We further note that for achieving a high degree of accuracy in transmission it is only necessary to ensure that the probability of erroneous decoding of the received code word remains small for all code words of 2^{HN} 'probable' N -letter messages (since all the remaining N -letter messages are very rarely encountered and the errors in their restoration make little impact).

We seek a coding method for which the length of a code word of N -letter blocks is $(H/c_1)N = N_1$ elementary signals;† here c_1 is a number chosen beforehand, which *must satisfy the unique condition*

$$c_1 < c$$

(but c_1 may be arbitrarily close to c !). The number of all distinct sequences of $(H/c_1)N$ elementary signals obviously equals $m^{(H/c_1)N} = 2^{(\log m/c_1)HN}$. Since $c_1 < c \leq H(\beta) \leq \log m$, it is certainly larger than 2^{HN} and hence a distinct sequence of $N_1 = (H/c_1)N$ elementary signals can be associated as a code word to each of 2^{HN} 'probable' N -letter messages. However, it is further required that the probability of erroneous decoding of all transmitted code words remain small. This clearly prescribes that the 2^{HN} code words used by us must differ sharply from each other, for only subject to such restriction can it be expected that in spite of the possible noise distortion of transmitted signals, it will be possible to distinguish the code words from one another at the channel output with sufficient reliability.

In order to estimate the possible number of such N_1 -term code words that are *effectively distinguishable from one another*, we may argue as follows. Every sequence of $N_1 = (H/c_1)N$ transmitted elementary signals A_i (where $i = 1, 2, \dots, m$) is received at the channel output as a chain of certain N_1 elementary signals B_j (where $j = 1, 2, \dots, r$; see p. 261). Obviously, by transmitting *one and the same* N_1 -sequence $A_{i_1}A_{i_2} \dots A_{i_{N_1}}$ many times we obtain at the output *many different*

†As is usual, if the number $(H/c_1)N = N_1$ is not an integer, then it is necessary to replace it by an *integer* closest to it. This remark relates also to all other numbers encountered below, which in their own right must necessarily be integers.

sequences $B_{j_1}B_{j_2} \dots B_{j_{N_1}}$. This fact reflects a random character of noise which affects the transmission. However, by transmitting a single N_1 -sequence $A_{i_1}A_{i_2} \dots A_{i_{N_1}}$ we shall obtain at the output different sequences $B_{j_1}B_{j_2} \dots B_{j_{N_1}}$ with different frequencies: one of them appears here relatively more frequently, others quite rarely, however.† The following arguments enable us to evaluate approximately the number of chains $B_{j_1}B_{j_2} \dots B_{j_{N_1}}$ which, with not too low probability, may arise in the transmission of the given chain $A_{i_1}A_{i_2} \dots A_{i_{N_1}}$. Assume that elementary signals A_i are successively transmitted through the communication channel, each time choosing a transmitted signal at random (and independent of all signals transmitted previously), with probabilities $p^0(A_1), p^0(A_2), \dots, p^0(A_m)$. In such a case, by what is stated on p. 168, for large N_1 , among all N_1 -signal chains of the form $A_{i_1}A_{i_2} \dots A_{i_{N_1}}$ only $2^{H(\beta)N_1}$ chains are 'probable' (all having nearly the same probability), while the probability sum of the transmission of one of the remaining chains (whose number is equal to $m^{N_1} - 2^{H(\beta)N_1} = 2^{\log m \times N_1} - 2^{H(\beta)N_1}$) is found to be very small. We agree to choose all N_1 -signal code words needed by us from $2^{H(\beta)N_1}$ 'probable' N_1 -signal chains and ignore completely the remaining such chains. This is possible since

$$H(\beta)N_1 = \frac{H(\beta)}{c_1} HN > HN$$

(because $c_1 < c \leq H(\beta)$) and, consequently, the total number of 'probable' chains also exceeds the number 2^{HN} of required code words.

Consider now all possible chains of the form $A_{i_1}A_{i_2} \dots A_{i_{N_1}} B_{j_1}B_{j_2} \dots B_{j_{N_1}}$ formed of N_1 transmitted elementary signals A_i and those N_1 signals B_j into which these signals A_i get converted after transmission through the communication channel. The total number of such $2N_1$ -term chains is obviously given by

$$m^{N_1}r^{N_1} = 2^{(\log m + \log r)N_1}.$$

The arguments adduced on p. 168 can be applied to these chains also, implying the following result: if all A_i are so chosen as explained above, then only $2^{H(\alpha\beta)N_1}$ are 'probable' out of the total number $2^{(\log m + \log r)N_1}$ of our chains. Moreover, the probabilities of all probable chains are nearly equal to each other, while the total probability of all the remaining $2^{(\log m + \log r)N_1} - 2^{H(\alpha\beta)N_1}$ chains

†For example, let us consider the case of the binary symmetric channel (see pp. 263-264). If the N_1 -sequence $A_{i_1}A_{i_2} \dots A_{i_{N_1}}$ is transmitted through such a channel, where N_1 is large enough, we obtain at the output obviously with a fairly large probability one of the N_1 -sequences distinct from the transmitted chain of signals in not less than $N_1(p - \delta)$ signals but in not more than $N_1(p + \delta)$ signals, where δ is some small number (see discussion of the law of large numbers in Section 1.4).

is quite small.[†] Consequently, the number of 'probable' $2N_1$ -term chains $A_{i_1}A_{i_2} \dots A_{i_{N_1}} B_{j_1}B_{j_2} \dots B_{j_{N_1}}$ exceeds the number of 'probable' N_1 -term transmitted chains $A_{i_1}A_{i_2} \dots A_{i_{N_1}}$ by

$$2^{H(\alpha\beta)N_1} : 2^{H(\beta)N_1} = 2^{[H(\alpha\beta)-H(\beta)]N_1} = 2^{H_{\beta}(\alpha)N_1}$$

times. Hence it can be concluded that a whole group of $2^{H_{\beta}(\alpha)N_1}$ chains $B_{j_1}B_{j_2} \dots B_{j_{N_1}}$ of received signals corresponds to every 'probable' N_1 -term transmitted chain $A_{i_1}A_{i_2} \dots A_{i_{N_1}}$, which gets converted into one of the chains $B_{j_1}B_{j_2} \dots B_{j_{N_1}}$ with a large probability (i.e., with a probability very close to

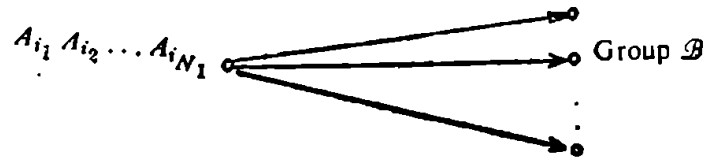


Fig. 25.

unity). For brevity, we designate this group of $2^{H_{\beta}(\alpha)N_1}$ chains $B_{j_1}B_{j_2} \dots B_{j_{N_1}}$ corresponding to $A_{i_1}A_{i_2} \dots A_{i_{N_1}}$, as *group B, corresponding to $A_{i_1}A_{i_2} \dots A_{i_{N_1}}$* (see the schematic diagram in Fig. 25). Combining each of $2^{H(\beta)N_1}$ chains $A_{i_1}A_{i_2} \dots A_{i_{N_1}}$ that are sufficiently 'probable' to be transmitted with $2^{H_{\beta}(\alpha)N_1}$ chains of the group B corresponding to it, we obtain precisely all $2^{H(\alpha\beta)N_1}$ 'probable' chains $A_{i_1}A_{i_2} \dots A_{i_{N_1}} B_{j_1}B_{j_2} \dots B_{j_{N_1}}$.

Two N_1 -term chains of transmitted signals $A_{i_1}A_{i_2} \dots A_{i_{N_1}}$ and $A'_{i_1}A'_{i_2} \dots A'_{i_{N_1}}$ should be considered to be 'effectively distinguished from each other' if two groups B corresponding to them are *disjoint*. In fact, the message $A_{i_1}A_{i_2} \dots A_{i_{N_1}}$ (respectively $A'_{i_1}A'_{i_2} \dots A'_{i_{N_1}}$) after transmission through our communication channel is 'almost surely' (i.e., with probability close to unity) transformed into one of the chains $B_{j_1}B_{j_2} \dots B_{j_{N_1}}$ belonging to the first (resp. second) group B. Hence, if the indicated two groups B are disjoint and it is known that either the message $A_{i_1}A_{i_2} \dots A_{i_{N_1}}$ or $A'_{i_1}A'_{i_2} \dots A'_{i_{N_1}}$ was transmitted, then we can, for instance, in all cases when one of the chains of the first group B is obtained at the output, assume that the message $A_{i_1}A_{i_2} \dots A_{i_{N_1}}$ was transmitted, but when any other chain is obtained (including also all chains of the second group B) it can be considered that $A'_{i_1}A'_{i_2} \dots A'_{i_{N_1}}$ was transmitted. Here it is clear that the probability of erroneous decoding of the received message will be fairly small. In analogy to this, if it is required to choose 2^{HN} different code words, each of them consisting of N_1 signals A_i , then in order that the probability of

[†]The result is based on the fact that any $2N_1$ -term chain $A_{i_1}A_{i_2} \dots A_{i_{N_1}} B_{j_1}B_{j_2} \dots B_{j_{N_1}}$ can be considered as a chain $(A_{i_1}B_{j_1})(A_{i_2}B_{j_2}) \dots (A_{i_{N_1}}B_{j_{N_1}})$ formed of N_1 successive outcomes of the compound experiment $\alpha\beta$ (with mr possible outcomes), having the entropy $H(\alpha\beta)$.

erroneous decoding of the received message be small, it suffices to have an opportunity to choose these code words in such a way that for all 2^{HN} groups the \mathcal{B} 's corresponding to them are disjoint. Since each group \mathcal{B} contains $2^{H_{\beta}(\alpha)N_1} = 2^{(H_{\beta}(\alpha)/c_1)HN}$ chains $B_{j_1}B_{j_2} \dots B_{j_{N_1}}$, there are

$$2^{\frac{H_{\beta}(\alpha)}{c_1}HN} \times 2^{HN} = 2^{\left(\frac{H_{\beta}(\alpha)}{c_1} + 1\right)HN}$$

chains in 2^{HN} groups \mathcal{B} . Moreover, since all such chains $B_{j_1}B_{j_2} \dots B_{j_{N_1}}$ terminate the 'probable' $2N_1$ -term sequences $A_{i_1}A_{i_2} \dots A_{i_{N_1}} B_{j_1}B_{j_2} \dots B_{j_{N_1}}$, they themselves are also 'probable', i.e., they belong to a set of N_1 -term chains of signals B_j that arise not too infrequently at the channel output when N_1 signals A_i are successively transmitted through channel and each time a transmitted signal is chosen at random with probabilities $p^0(A_1), p^0(A_2), \dots, p^0(A_m)$ (regardless of what signals were transmitted earlier). The number of such 'probable' chains $B_{j_1}B_{j_2} \dots B_{j_{N_1}}$ (i.e., the 'probable' chains of N_1 successive outcomes of experiment α), as is known, is given by

$$2^{H(\alpha)N_1} = 2^{\frac{H(\alpha)}{c_1}HN}.$$

Let us now construct the ratio of the total number $2^{(H(\alpha)/c_1)HN}$ of 'probable' chains $B_{j_1}B_{j_2} \dots B_{j_{N_1}}$ to the total number $2^{[(H_{\beta}(\alpha)/c_1)+1]HN}$ of such chains appearing in 2^{HN} groups \mathcal{B} :

$$\begin{aligned} \frac{2^{\frac{H(\alpha)}{c_1}HN}}{2^{\left(\frac{H_{\beta}(\alpha)}{c_1} + 1\right)HN}} &= 2^{\left(\frac{H(\alpha)}{c_1} - \frac{H_{\beta}(\alpha)}{c_1} - 1\right)HN} = 2^{\left(\frac{H(\alpha) - H_{\beta}(\alpha)}{c_1} - 1\right)HN} \\ &= 2^{\left(\frac{c}{c_1} - 1\right)HN}. \end{aligned}$$

It is seen that if c_1 is *larger* than c , then this ratio is less than unity, i.e., the total number of chains in our 2^{HN} groups \mathcal{B} is larger than the total number of all 'probable' chains $B_{j_1}B_{j_2} \dots B_{j_{N_1}}$. It is hence clear that for $c_1 > c$ it is *impossible* to select in any way whatsoever the 2^{HN} code words such that all groups \mathcal{B} corresponding to them are disjoint. This obviously is as it should be, since it is already known to us that, if $c_1 > c$, then it is impossible to transmit through our channel a message at a rate of $L(c_1/H)$ letters per unit time, and to achieve an arbitrarily small probability of its erroneous decoding at the channel output. But if c_1 is *less* than c , then the ratio set forth above is found to be greater than unity (since in this case $(c/c_1) - 1 > 0$); furthermore, for extremely large N it is found to be equal to 2 raised to a large power, i.e., *exceedingly large*. Thus, for large N , the total number of chains in 2^{HN} groups \mathcal{B} form a *negligible share* of the overall number of 'probable' chains from N_1 signals B_j . The last situation

yields a highly plausible premise that 2^{HN} code words of length $(H/c_1)N$ can be chosen such that all groups \mathcal{B} corresponding to them are disjoint. Moreover, as we know, such a choice of code words assures for large N the possibility of deciphering the obtained message with an arbitrarily small probability of error.

The arguments set forth above lend great plausibility to Shannon's theorem, but obviously we cannot regard them as a mathematical proof of this theorem (this situation is more elaborately explained on pp. 290–291). In spite of this, for the present we confine ourselves to all that has been stated and pass on to analyze certain other problems connected with Shannon's theorem. Later, however, we shall adduce on pp. 291–297, following Shannon, interesting (but not quite simple) arguments to prove conclusively that indeed there must exist such a choice of 2^{HN} code words that guarantees, if not complete non-overlapping of corresponding 2^{HN} groups \mathcal{B} , then at least guarantees that this overlapping be sufficiently small so as not to affect the fact that the probability of erroneous decoding can be made arbitrarily close to zero. At the end of the present section (pp. 298–303) we shall analyze in greater depth another rigorous proof of the fundamental coding theorem, though related only to the special case of a binary symmetric channel. It is left to the reader to decide whether it is worthwhile to devote his time to all this material (and when, whether now or later, he should follow the plan of exposition as given in the book), or whether he should prefer to confine himself to just the non-rigorous reasonings set forth above; in the latter case the entire concluding portion of this section (from pp. 290 to 304) may be skipped by the reader.

The reader may only be cautioned in advance that both proofs of Shannon's theorem exposed at the end of the section are *noneffective* in the same way as are all other known proofs of it: from them it follows that for sufficiently large N there necessarily exists such a method of choice of code words as to guarantee that the probability of error in reconstruction of each letter of the obtained message does not exceed a given (arbitrarily small) number ϵ , but nothing is said about how one can find such a method of choice of code words (see also the beginning of the next section, where this position has been explained more precisely). The question as to how in fact the code words ought to be chosen in order that the probability of error in decoding be made sufficiently small is dealt with in the next and last section of the book.

It was noted above that Shannon's theorem does not allow us to indicate in what specific manner we ought to choose code words in order that a message be transmitted through a communication channel at a given rate

$$v_1 < v = L \frac{c}{H} \text{ letters per unit time,}$$

and also in order that the corresponding probability of transmission error does not exceed a given small number ϵ . Let us now note that this theorem does not permit us also to state how large the number N of letters in coded blocks must be in order that such transmission become possible. This theorem implies only that if it is permitted to choose N *arbitrarily large*, then transmission with speed v_1 and error probability not exceeding ϵ are *possible*, whatever $v_1 < v$ and $\epsilon > 0$. However, since with increasing N the complexity of deciphering a code is considerably increased and further time-lag in deciphering is involved, it is of practical interest to be able to evaluate also the least value of error probability ϵ attainable in transmission at a given speed v_1 by means of a code whose code words correspond to letter blocks *consisting of not more than N letters*, where N is some *given* number. This problem has been dealt with by C. E. Shannon, A. Feinstein, P. Elias, J. Wolfowitz, R. G. Gallager, R. L. Dobrushin, and other scientists; a detailed exposition and proof of the results obtained by them can be found, for instance, in the papers [181]-[183], [186] and books [2], [8], [9], [11] and [23], which are quite complex. Without going into details we shall simply state here the basic fact stemming from all these investigations.

Recall that the transmission of an N -letter block at a rate $v_1 = L(c_1/H)$ letters per unit time, where $c_1 < c$, is attained when we assign to individual N -letter blocks, code words consisting of $N_1 = (H/c_1)N$ elementary signals. It is appropriate to use c_1 and N_1 in place of v_1 and N when calculating the error probability corresponding to the given values of $v_1 = L(c_1/H)$ and N , since c_1 and N_1 describe more directly the process of information transmission over the communication channel. It is found that *for fixed $c_1 < c$ and N_1 there always exists a method of transmission* (i.e., a coding method that permits us to choose $2^{c_1 N_1}$ code words consisting of N_1 elementary signals, and a decoding method that gives the rule for deciphering the received N_1 -term chains of elementary signals B_1), *for which the probability of erroneous interpretation of every transmitted code word does not exceed the quantity*

$$\epsilon = \frac{1}{a^{N_1}},$$

where a is some number greater than unity.[†] The number a obviously depends on c_1 , the smaller the c_1 (i.e., factually smaller the rate v_1 of information trans-

[†]The formula derived here can of course be rewritten as $\epsilon = 1/a_1^{N_1}$, where $a_1 = aH/c_1$ is a new number (this also is greater than unity). However, a_1 is found to depend also on the entropy H of the transmitted message, whereas a depends only on the value of c_1 and the characteristics of the communication channel used. For the reader acquainted with natural logarithms, it is useful to keep in view also that in the scientific literature the formula for ϵ is usually written in the form $\epsilon = e^{-EN_1}$, where $e = 2.718 \dots$ is the base of natural logarithms and $E = \ln a$ is the natural logarithm (with base e) of a . Since $y = e^{-Ex}$ is the so-called exponential function, the preceding formula for ϵ is frequently called the *exponential bound of error probability*, or simply the *exponential bound of error*.

mission through the communication channel), the larger is a . It seems natural to expect that when c_1 (and, hence also v_1) tends to zero the number a increases indefinitely (since by decreasing indefinitely the information transmission rate, an arbitrarily small error probability can be attained for any fixed N). Actually, however, all derivations of the above mentioned formula for ϵ for quite low transmission speeds are found to be rather crude and they usually indicate that a tends to a finite value as $c_1 \rightarrow 0$. When c_1 tends to c (i.e., the transmission rate v_1 to v), the number a tends to unity, so that ϵ also approaches unity with the growth of v_1 . The value of a for given c_1 is different for different communication channels; a schematic diagram of the dependence of a on c_1 for a fixed channel is given in Fig. 26.

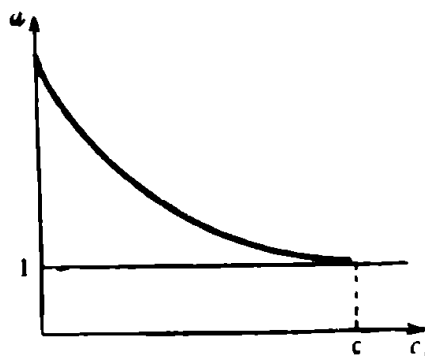


Fig. 26.

It is clear that Shannon's noisy coding theorem directly follows from the formula indicated for ϵ and the fact that $a > 1$ for any $c_1 < c$. Moreover, this formula extends appreciably Shannon's theorem, which says only that ϵ can be made arbitrarily small if only N (or, equivalently, N_1) is chosen sufficiently large (but states nothing about precisely how ϵ decreases with the growth of N). It is precisely the last situation we had in our view on p. 275 when we remarked that the results obtained in [186] are sharper than the fundamental coding theorem.

We now pass on to the case of message transmission at a rate v_1 *greater* than the limit rate $v = L(c/H)$ letters per unit time. This is in general of less interest than the case of transmission at a rate $v_1 < v$ and the results related to it are less spectacular than Shannon's fundamental theorem; nevertheless it merits our examination. We have already noted on p. 273 that an *error-free* information transmission cannot take place at a rate $v_1 > v$ letters per unit time; a similar statement may also be found on p. 279 where it is indicated that if $c_1 > c$, then 2^{HN} groups \mathcal{B} , corresponding to the code words of all possible 'probable' N -letter blocks, can in no way be so chosen that they are disjoint. In reality, however, the reasonings given on pp. 273 and 279 allow us to draw

only superficial conclusions. It is of course true that an error-free message transmission cannot be accomplished at a rate exceeding $v = C/H$ letters per unit time. However, as a matter of fact, even in the case of transmission at a rate $v_1 < v$ we cannot assert that error-free message transmission is possible but can state only that in this case the probability of erroneous interpretation of every transmitted letter can be made as small as desired (by using sufficiently long chains of elementary signals as code words).† Hence, a precise statement of the converse to Shannon's fundamental noisy coding theorem must not assert that for $v_1 > v$ an *error-free* information transmission is impossible, but rather that *for any fixed $v_1 > v$ there can be found a positive number $q_0 > 0$ (which apparently must depend on v_1 and increase for increasing v_1) such that in the case of information transmission through a communication channel at a rate v_1 the probability q of erroneously deciphering every transmitted letter of the message for any method of coding and decoding (independent of the values of N and N_1) is not less than q_0* . A conjecture on the validity of such a converse to the noisy coding theorem was also made by Shannon [21] and later it was rigorously proved by Fano [8]; we shall now proceed to consider its proof following Fano.

In the first place, however, the very statement of the theorem under consideration needs some sharpening. It is easy to see that the statement made about the probability of erroneously decoding cannot be necessarily true for all transmitted letters. Indeed, we can, for instance, stipulate that we shall decipher all received letters as the first letter of the alphabet—here the error probability will be zero in all cases in which the first letter is actually transmitted. On the other hand, it is also clear that to decipher all received letters as the first letter is inappropriate—here, in fact, we generally make no use whatsoever of the communication channel and commit an error every time a letter other than the first letter is transmitted; hence the mean error probability in this case will be large. At the same time it is most natural to understand the probability q of erroneously decoding a single transmitted letter specifically as the *mean error probability* and hereafter we shall indeed do so.

Thus, assume that the transmitted text is written by means of an n -letter alphabet a_1, a_2, \dots, a_n , and that the probabilities of the appearance of letters a_1, a_2, \dots, a_n at arbitrary (but fixed) places in this text are respectively equal to the given numbers p_1, p_2, \dots, p_n . By q we understand the *mean value of*

†It may be noted in this connection that Shannon [185] had introduced also the concept of the *zero error capacity C_0 of the channel*, defining it as the *highest rate* (in bits per unit time) at which completely *error-free* information transmission can be conducted over a given communication channel. The reasoning on p. 273 shows only that, no matter what communication channel, C_0 cannot exceed the channel capacity C defined on p. 263, a situation that seems to be almost obvious. In fact, the zero error channel capacity C_0 is usually appreciably smaller than C ; curiously, C_0 is found to be a more complex quantity than the usual channel capacity C , the value of C_0 is generally considerably more difficult to evaluate and it has lesser intuitive content.

error probability, i.e., the quantity

$$q = p_1 q_1 + p_2 q_2 + \dots + p_n q_n, \quad (*)$$

where q_1 is the probability that the letter a_1 after transmission through the communication channel is erroneously understood at the output as an alphabet letter other than a_1 and the quantities q_2, \dots, q_n carry a similar sense. It is essential that we are able to calculate this mean value q differently also. Suppose that p'_1, p'_2, \dots, p'_n are the probabilities of finding the letters a_1, a_2, \dots, a_n at an arbitrary (but fixed) place of the message obtained at the channel output by deciphering a received sequence of elementary signals B_j . Furthermore, denote by q'_1 the probability that the letter a_1 is obtained at the output due to incorrect deciphering of the received message (i.e., the corresponding place of the transmitted message is in fact occupied by a letter other than a_1), and by q'_2, \dots, q'_n similar probability errors related to the cases of having obtained the letters a_2, \dots, a_n . It is clear that the probabilities p'_1, p'_2, \dots, p'_n in general do not coincide with p_1, p_2, \dots, p_n (they depend on the probabilities p_1, p_2, \dots, p_n , and also on the coding-decoding method and the characteristics of the communication channel). However, the mean value of error probability for a single received letter can be expressed also in terms of p'_1, p'_2, \dots, p'_n , namely†

$$q = p'_1 q'_1 + p'_2 q'_2 + \dots + p'_n q'_n. \quad (**)$$

It is precisely the formula (**) that we shall use primarily hereafter.

Taking up the proof of the converse to the noisy coding theorem, we start with the simple case in which the transmitted message is written by means of a two-letter alphabet (for convenience, we denote by a and b the alphabet letters in this case). Suppose that β is an experiment consisting of determining at the input an alphabet letter of the message transmitted through the communication channel (not an elementary signal, as on pp. 261–262, but exactly a letter!), and that α is another experiment consisting of deciphering a letter at the channel output. Then, both these experiments can have two outcomes (a and b), the probabilities of the two possible outcomes of α being p'_1 and p'_2 (so that $p'_1 + p'_2 = 1$), and those of β given that α has the outcome a (resp. b) being $1 - q'_1$ and q'_1 (resp. q'_2 and $1 - q'_2$). Consequently,

$$H_a(\beta) = -q'_1 \log q'_1 - (1 - q'_1) \log (1 - q'_1) = h(q'_1),$$

$$H_b(\beta) = -q'_2 \log q'_2 - (1 - q'_2) \log (1 - q'_2) = h(q'_2),$$

†It is not difficult to understand that the right-hand sides of both equations (*) and (**) define the *mean frequency of errors* in successive deciphering of a large number of letters of the transmitted message.

where $H_a(\beta)$ and $H_b(\beta)$ are the conditional entropies of β given that α has the outcomes a and b , respectively, and, as usual,

$$h(p) = -p \log p - (1 - p) \log (1 - p).$$

Using the relations,

$$H_a(\beta) = h(q'_1), \quad H_b(\beta) = h(q'_2),$$

we obtain

$$H_\alpha(\beta) = p'_1 H_a(\beta) + p'_2 H_b(\beta) = p'_1 h(q'_1) + p'_2 h(q'_2).$$

We now make use of the fact that $h(p)$ (whose graph is given in Fig. 8 on p. 49) is a convex function in the sense explained in Appendix I on p. 347. Hence, by Theorem 2 of Appendix I (p. 350) for any nonnegative p'_1 and p'_2 such that $p'_1 + p'_2 = 1$, we have

$$p'_1 h(q'_1) + p'_2 h(q'_2) \leq h(p'_1 q'_1 + p'_2 q'_2) = h(q),$$

where $q = p'_1 q'_1 + p'_2 q'_2$. Thus

$$H_\alpha(\beta) \leq h(q), \tag{A}$$

and

$$I(\alpha, \beta) = H(\beta) - H_\alpha(\beta) \geq H(\beta) - h(q).$$

We now recall that $I(\alpha, \beta)$ is the information contained in an arbitrary text letter obtained at the channel output, concerning the corresponding letter of the transmitted message. Through the channel v_1 letters are transmitted per unit time, i.e., the amount of information transmitted per unit time equals $v_1 I(\alpha, \beta)$ (the successive letters of the message are considered to be mutually independent). But the amount of information transmitted per unit time cannot exceed the channel capacity C of our channel;† hence, furthermore,

$$v_1 [H(\beta) - h(q)] \leq C.$$

†Recall that C is the maximum information about the transmitted *elementary signals* that can be extracted from the elementary signals obtained per unit time at the output. If the encoding of a sequence of letters of a message into a sequence of elementary signals is not unique (for example, if the random coding described below on p. 292 is used), then the passage from α to experiment α_1 , consisting of determining the transmitted signal, is accompanied with some loss of information; nonunique decoding will also have a similar effect. For us here, however, the only important fact is that in every case the information $v_1 I(\alpha, \beta)$ about the transmitted letters contained in the received letters cannot be greater than C (see p. 89).

Since $C/H(\beta) = v$, the preceding inequality can be conveniently written in the form

$$1 - \frac{h(q)}{H(\beta)} \leq \frac{v}{v_1}. \quad (\text{B})$$

Consider the graph of the function $1 - (h(q)/H(\beta)) = g(q)$ (see Fig. 27a, b in which this function is depicted for the case in which $H(\beta) = 1$, i.e., when the

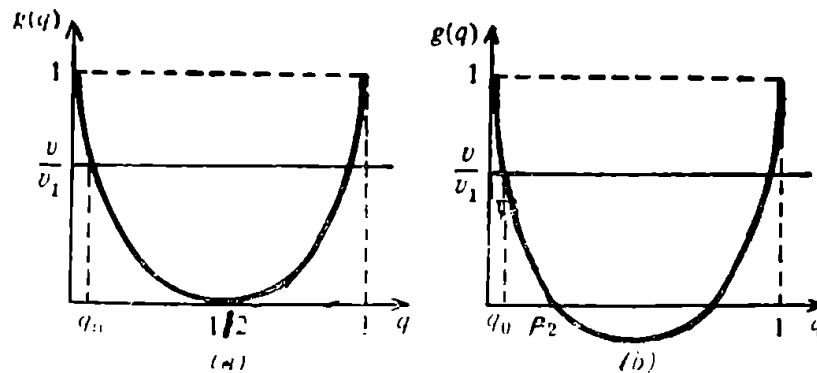


Fig. 27.

outcomes a and b of β are equally probable, and for the case in which $H(\beta) < 1$). The graph shows that if $v_1 \leq v$, i.e., if $v/v_1 \geq 1$, then inequality (B) can be satisfied for all values of q , including $q = 0$. If, however, $v_1 > v$, i.e., $v/v_1 < 1$, then this inequality can be fulfilled if and only if the value of q belongs to some interval of values lying to the left of the point q_0 , where $q_0 > 0$.

Thus, for $v_1 > v$ the mean error probability q cannot be less than a certain $q_0 > 0$, i.e., we have proved the statement designated above as the converse to the noisy coding theorem. With the growth of v_1 (i.e., with decreasing v/v_1) the value of q_0 increases; as $v_1 \rightarrow \infty$ (i.e., $v/v_1 \rightarrow 0$), q_0 obviously tends to the probability p_2 of the transmission of that one of the letters a or b which is transmitted *less frequently* than the other letter. The last result is clearly quite natural; in fact, when the transmission rate is extremely large we can transmit almost no useful information through our channel, and hence the most reasonable deciphering method in this case is the one by which *all* accepted letters are deciphered as a letter having the highest probability of being transmitted. But for such deciphering the mean error probability q is obviously equal to the probability of a more infrequent letter among the used letters a and b (note that for the indicated 'deciphering' a communication channel is not needed at all). If, however, the probability of the appearance of both text letters is the same, then for an extremely large transmission rate, when a communication channel is generally found to be of no practical use, there is no basis at all for us to choose this or the other value of the received letter, so that deciphering can be carried out here completely 'at random'. The mean error probability q in this

case as $v_1 \rightarrow \infty$ tends to $\frac{1}{2}$, since this is also the probability of erroneous decoding 'at random' (and simultaneously the probability of a 'more infrequent')

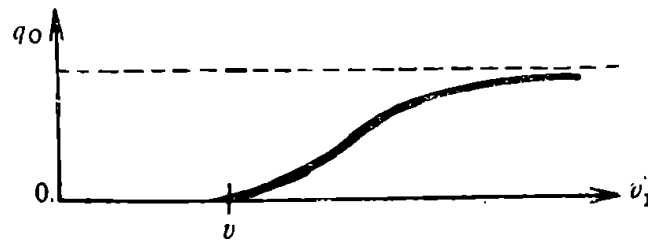


Fig. 28.

letter). A schematic graph of the dependence of the lower bound q_0 of error probability on the transmission rate v_1 is given in Fig. 28. The coincidence of the graph for $v_1 < v$ with the abscissa (i.e., $q_0 = 0$) obviously corresponds to Shannon's fundamental coding theorem which asserts that for $v_1 < v$ the error probability can be made as small as desired. (However, our conclusion proving only that the mean probability error *cannot be smaller than* q_0 for fixed value $v_1 > v$ does not by itself yield the assertion that for $v_1 < v$ the quantity q can indeed be made as small as desired.) The positiveness of q_0 for all $v_1 > v$ just forms the content of the converse to the coding theorem.

The case in which the transmitted message is written in a language employing an alphabet of n letters a_1, a_2, \dots, a_n is not considerably more complicated than the particular case of the two-letter alphabet analyzed above. Here, however, in place of the completely elementary inequality (A) we have to make use of a more general *Fano inequality* having the form

$$H_\alpha(\beta) \leq h(q) + q \log(n-1), \quad (A')$$

where α and β have the same meaning as above and q is again the mean error probability.

Fano's inequality (A') has a quite simple and intuitive meaning. In fact, $H_\alpha(\beta)$ is the mean amount of uncertainty of the outcome of β when the outcome of α is known. But the outcome of β given the outcome of α can be ascertained by means of the following two auxiliary experiments. First we determine *whether or not the error would occur in the transmission of the corresponding letter of the message*. This implies that we carry out an experiment γ capable of having only two outcomes (answers 'yes: it occurs', or 'no: it does not occur'). The mean probability of the outcome of γ being positive (answer 'yes') obviously equals q . Making use of the convexity of the function $h(p)$ it is, therefore, easy to infer that the mean amount of uncertainty of the result of our first auxiliary experiment cannot exceed $h(q)$ (see on p. 285 the inequality preceding (A) and also the similar general derivation on p. 304). It is further clear that if the error in the transmission does not occur (i.e., if the outcome of γ is negative),

then the results of γ and α uniquely determine the outcome of β . If, however, the outcome of γ is found to be positive (which happens on an average in a portion q of all cases), then the knowledge of the outcome of γ does not remove all uncertainty in the outcome of β , necessitating an extra auxiliary experiment γ_1 in order to ascertain exactly what letter out of $n - 1$ letters other than those received was indeed transmitted. Experiment γ_1 can have $n - 1$ different outcomes; hence the amount of its uncertainty (the entropy of γ_1) cannot exceed $\log(n - 1)$. It is clear that the total amount of uncertainty $H(\beta)$ must be equal to the amount of uncertainty of the first auxiliary experiment γ added to the amount of uncertainty of the second experiment γ_1 , multiplied by the mean frequency of the cases in which γ_1 is found to be needed. This immediately implies Fano's inequality (A') (for more details on this, see the text in small print on pp. 303-304).

We now note that Fano's inequality implies the inequality

$$I(\alpha, \beta) \geq H(\beta) - h(q) - q \log(n - 1).$$

Hence

$$v_1[H(\beta) - h(q) - q \log(n - 1)] \leq C,$$

where $C = vH(\beta)$, i.e.,

$$1 - \frac{h(q) - q \log(n - 1)}{H(\beta)} \leq \frac{v}{v_1}. \quad (B')$$

In the particular case in which $H(\beta) = \log n$, the function

$$g_n(q) = 1 - \frac{h(q) - q \log(n - 1)}{H(\beta)}$$

differs from the function $C(p)$ depicted in Fig. 20 on p. 267 (for the particular case $n = 4$) only by a constant factor; for convenience we draw a similar graph (Fig. 29a). Schematic forms of the graph of the function $g_n(q)$ for $H(\beta) < \log n$ (i.e., when not all alphabet letters are equally probable) are given side by side in Fig. 29b. We see that if $v_1 < v$ (i.e., $v/v_1 > 1$), then inequality (B') holds for any $q \geq 0$; if, however, $v_1 > v$ (i.e., $v/v_1 < 1$), then (B') is satisfied only for values of q larger than some positive number q_0 . This shows the converse to the coding theorem to be true also in the general case of n -letter alphabets. The dependence of the value of q_0 on the transmission rate v_1 here again has the form schematically depicted in Fig. 28; the limiting value of q_0 as $v_1 \rightarrow \infty$ (i.e., as $v/v_1 \rightarrow 0$) in the case in which $H(\beta) = \log n$ is equal to $(n - 1)/n$ and it decreases with decreasing $H(\beta)$.†

†If v_1 is quite large, then the communication channel becomes useless and hence here it remains to decode all the received letters as the letter having the highest probability of being transmitted. In this case, the mean error probability q equals $1 - p_1$, where p_1 is the largest of the probabilities of alphabet letters. Since, however, the inequality (B') is not exact, an estimate obtained from it of the lower bound q_0 of the mean error probability will not in general necessarily coincide with the lowest actually attainable value of q .

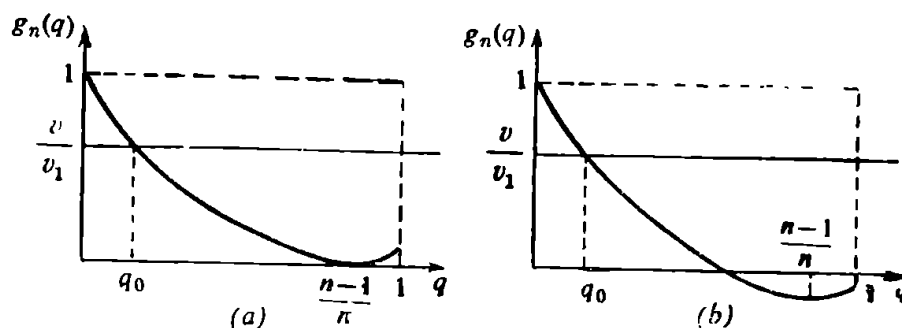


Fig. 29.

Let us note that by what has been proved in the present section the *fundamental noisy coding theorem* and the *converse to the noisy coding theorem* differ sharply both in the method of proof and in their character as well. It is true that the probability of erroneous determination of a single transmitted *letter* appears in the statement of both theorems. In fact, however, while considering the fundamental coding theorem, the original message in letters was just briefly touched upon at the start of our discussion and then we concentrated only on chains of N_1 *elementary signals* transmitted directly through the communication channel. The essential fact here was just the following: if we use code words ('blocks') consisting of N_1 elementary signals, then transmission at a rate $v_1 = L(c_1/H)$ letters per unit time demanded that these code words correspond to N -letter messages, where $N = (c_1/H)N_1$, i.e., that (in the case of N_1 sufficiently large) there are not less than $2^{HN} = 2^{c_1 N_1}$ 'probable' code words. Thus, it was required to show only that, if $c_1 < c$ (where $c = \max I(\alpha, \beta)$), then for sufficiently large N_1 it is always possible to choose $2^{c_1 N_1}$ code words of length N_1 in such a way that the probability of erroneous decoding of a chain of N_1 elementary signals obtained at the channel output will be less than an arbitrary (but preassigned) number ϵ , irrespective of the specific code word that is being transmitted (here naturally ϵ is chosen very small, say, equal to 0.001, or 0.0001, or 0.000001). This statement (related just to a communication channel and lengthy chains of *elementary signals* transmitted through it, but by no means connected to the original message in letters) forms exactly the essence of the fundamental coding theorem. As regards the converse to the coding theorem, it is essentially related to the *letters* of the message and not to chains of elementary signals transmitted through the communication channel.

There exists another theorem which also is the converse to the fundamental coding theorem, but is concerned only with a communication channel and lengthy chains of *elementary signals* transmitted through it. This theorem says that, if $c_1 > c$ and N_1 is sufficiently large, then no matter what $2^{c_1 N_1}$ equally probable code words (i.e., chains of elementary signals) of length N_1 are chosen and what method of deciphering the received N_1 -term chains of elementary signals is used, the mean probability of erroneous decoding of the received chain all the same exceeds an arbitrary (but preassigned) number $p_0 < 1$ (the number p_0 is

naturally chosen here sufficiently close to unity, say, equal to 0.999, or 0.9999, or 0.999999). It is, of course, clear that the closer p_0 is to unity, the larger is the required value of N_1 . As regards the mean error probability in the statement of the theorem, it obviously coincides with the *arithmetic mean*

$$\frac{p_{0,1} + p_{0,2} + \dots + p_{0,2^{c_1 N_1}}}{2^{c_1 N_1}}$$

where $p_{0,i}$ is the probability of a decoding error when the i th of our $2^{c_1 N_1}$ code words is transmitted.

The validity of the stated theorem is closely related to the discussion on p. 279. It was shown there that for $c_1 > c$ and very large N_1 the total number of N_1 -term chains in $2^{c_1 N_1}$ groups \mathcal{B} (i.e., in groups of received 'probable' chains corresponding to the $2^{c_1 N_1}$ 'probable' code words of length N_1) greatly exceed the total number of all 'probable' received chains. Hence the N_1 -term chains received belong in general simultaneously to the vast number of different groups \mathcal{B} , so that the probability of their correct decoding is quite low. These arguments lend utmost credibility to our theorem, even though they cannot be a substitute for its rigorous proof. Such proof can be found, for instance in [2], [11] or [23]; this proof is not quite straightforward and we shall not dwell upon it here. The theorem under consideration itself is called by Wolfowitz (the first to prove it rigorously) the *strong converse of the noisy coding theorem* and it is frequently referred to by this designation in the literature on information theory. However, the designation is not quite appropriate, since it may create a wrong impression that the conventional converse of the noisy coding theorem proved above follows from this new theorem (in fact, neither of the two converse theorems derived here is a consequence of the other). Hence, apparently, it shall be more appropriate that following Gallager [11] the theorem under consideration is called the *converse of the noisy block coding theorem* (i.e., the coding theorem which uses as code words blocks of elementary signals of a fixed length).

We now pass on to the more accurate proof of Shannon's fundamental noisy coding theorem of which we spoke on p. 274 et seq. To start with, following Zaremba [188], we give an example which shows quite clearly that from the fact of the total number of chains $B_{j_1} B_{j_2} \dots B_{j_{N_1}}$ in 2^{HN} groups \mathcal{B} being exceedingly small in comparison to the total number of such 'probable' chains, it still does not follow at all that these groups can be chosen to be such that they are disjoint. Consider from this objective a collection of all possible chains of 10 elementary signals, each of which can take two values. It is clear that the total number of such chains is $2^{10} = 1024$. We further associate with each chain of the group all 10-term chains differing from the given chain by not more than

three signals. Besides the given chain, this group obviously contains $\binom{10}{1} = 10$, $\binom{10}{2} = 45$ and $\binom{10}{3} = 120$ chains differing from the given chain exactly by one, two and three signals, respectively; therefore, the whole group consists of $1 + 10 + 45 + 120 = 176$ chains. Since 176 is very close to being $\frac{1}{8}$ of 1024, it might be thought that *three* chains could be chosen here without any singular difficulty such that three groups of 176 chains, corresponding to them, would be disjoint. But this is not correct: it can be shown that the *groups corresponding to any three chains necessarily intersect*.

Indeed, let us denote two values of our signals by the digits 0 and 1 and let, for instance, the first group correspond to a 'zero chain' of ten zeros. It is easy to understand that only the groups corresponding to 10-term chains containing more than six 1's do not intersect with the first group. But in every two 10-term chains each containing seven or more 1's, not less than four of these 1's would lie in both the chains at one and the same place. Consequently, the two given chains differ from each other by signals at not more than six places and hence the groups corresponding to them intersect with each other. Obviously, nothing is altered, if we start with any other chain (and not with the 'zero chain' 0000000000): our two groups of 176 chains not intersecting with one and the same third group necessarily intersect with each other.

In exactly the same way it can also be shown that *for any k among all groups of $(3k + 1)$ -term chains, differing from some one such chain by not more than k signals, it is impossible to find more than two disjoint groups*. Meanwhile, it can be easily verified that the ratio of the number of chains in such a group (equal to the sum $1 + \binom{3k+1}{1} + \binom{3k+1}{2} + \dots + \binom{3k+1}{k}$) to the total number of all possible $(3k + 1)$ -term chains ($= 2^{3k+1}$) will always decrease with increasing k . Thus, for $k = 8$, $3k + 1 = 25$ this ratio is close to $1/20$, and if k is chosen sufficiently large, the indicated ratio can be made *as small as desired* (smaller than any preassigned small number). Thus, the total number of chains in three groups comprises an insignificant part of the number of all possible chains, but nevertheless any three groups necessarily intersect. Hence, in the case of Shannon's theorem also, it is impossible to justify the possibility of choosing 2^{HN} disjoint groups by the fact that the total number of chains in them is very small in comparison to the number of all 'probable' chains. It has also to be proved rigorously that in a given case the situation is not such as that in the example due to Zaremba.

In fact, none has so far succeeded in proving rigorously that 2^{HN} chains $A_{i_1} A_{i_2} \dots A_{i_{N-1}}$ can be chosen in such a way that any two of the 2^{HN} groups \mathcal{B} corresponding to them are disjoint. However, it can be shown that there certainly exists a choice of these chains such that the corresponding groups \mathcal{B} are *almost disjoint* and hence their overlapping can be ignored. This fact can be made clear by means of the following arguments mainly due to Shannon [21].

To start with, we choose the requisite 2^{HN} chains $A_{i_1}A_{i_2} \dots A_{i_{N_1}}$ by a method which may seem, at first sight, to be clearly unreasonable, and specifically 'at random.' This choice 'at random' can be accomplished thus: we number all $2^{H(\beta)N_1}$ 'probable' chains $A_{i_1}A_{i_2} \dots A_{i_{N_1}}$ in an arbitrary order, write out their numbers on $2^{H(\beta)N_1}$ pieces of paper, place all these papers in an urn and mix them well, and then draw the pieces of paper one at a time from the urn 2^{HN} times, replacing the paper drawn after each draw and again mixing the contents of the urn. The chains $A_{i_1}A_{i_2} \dots A_{i_{N_1}}$ with the numbers drawn we take as our 2^{HN} code words (such a method of choosing code words is called *random coding*). It is clear that under random coding the same number may be drawn two or more times, so that some of the 2^{HN} selected chains turn out to be *identical* to each other and obviously they cannot be distinguished by any means at the receiving end; this situation alone gives an impression that the suggested method of choosing code words is undoubtedly irrational. In fact, however, for large N the probability of such coincidence is negligibly small (since the number $2^{H(\beta)N_1} = 2^{(H(\beta)/c_1)HN}$ of different 'probable' chains, when N is large, is many times larger than the number 2^{HN}). As we shall see later, this allows us to ignore completely the possibility of coincidence.

We now assume that the signals $A_{i_1}, A_{i_2}, \dots, A_{i_{N_1}}$ are transmitted successively through our communication channel, the collection of which forms precisely one of the code words chosen by us. Because of the presence of noise, these signals are in general somewhat distorted during transmission; as a result, we obtain at the receiving end of the channel a sequence of signals $B_{j_1}B_{j_2} \dots B_{j_{N_1}}$ different from $A_{i_1}A_{i_2} \dots A_{i_{N_1}}$. It is clear that the chain $B_{j_1}B_{j_2} \dots B_{j_{N_1}}$ belongs with probability quite close to unity to the group \mathcal{B} corresponding to the chain $A_{i_1}A_{i_2} \dots A_{i_{N_1}}$. But this chain $B_{j_1}B_{j_2} \dots B_{j_{N_1}}$ will at the same time belong also to groups \mathcal{B} corresponding to many other chains of N_1 transmitted signals. This specific circumstance makes it difficult to decipher the received message.

It is rather easy to estimate the total number of different 'probable' chains $A_{i_1}A_{i_2} \dots A_{i_{N_1}}$ having the property that the groups \mathcal{B} corresponding to them contain the given chain $B_{j_1}B_{j_2} \dots B_{j_{N_1}}$. In fact, the total number of 'probable' $2N_1$ -term chains $A_{i_1}A_{i_2} \dots A_{i_{N_1}} B_{j_1}B_{j_2} \dots B_{j_{N_1}}$, as we know, is $2^{H(\alpha\beta)N_1}$, and the chains $B_{j_1}B_{j_2} \dots B_{j_{N_1}}$ appearing in them all belong to the set of $2^{H(\alpha)N_1}$ equally likely 'probable' received chains. Thus, the number of 'probable' $2N_1$ -term chains exceeds the number of 'probable' N_1 -term chains $B_{j_1}B_{j_2} \dots B_{j_{N_1}}$ by $2^{H(\alpha\beta)N_1} : 2^{H(\alpha)N_1} = 2^{H_{\alpha}(\beta)N_1}$ times. It can hence be concluded that all possible 'probable' $2N_1$ -term chains are obtained by combining each of $2^{H(\alpha)N_1}$ 'probable' chains $B_{j_1}B_{j_2} \dots B_{j_{N_1}}$ of the received signals with $2^{H_{\alpha}(\beta)N_1}$ different chains $A_{i_1}A_{i_2} \dots A_{i_{N_1}}$ of transmitted signals. It is precisely these $2^{H_{\alpha}(\beta)N_1}$ transmitted chains that possess the property that the given chain $B_{j_1}B_{j_2} \dots B_{j_{N_1}}$ enters the groups \mathcal{B} corresponding to them. A collection of all these chains $A_{i_1}A_{i_2} \dots A_{i_{N_1}}$ we call *group \mathcal{A} corresponding to the chain $B_{j_1}B_{j_2} \dots B_{j_{N_1}}$* (see

the schematic Figure 30 in which the arrows from the chains of group \mathcal{A} to the chain $B_{j_1}B_{j_2} \dots B_{j_{N_1}}$ indicate that all groups \mathcal{B} of these transmitted chains contain $B_{j_1}B_{j_2} \dots B_{j_{N_1}}$ and that, consequently, there exists the real probability of any chain of group \mathcal{A} to be received at the channel output as the chain $B_{j_1}B_{j_2} \dots B_{j_{N_1}}$.

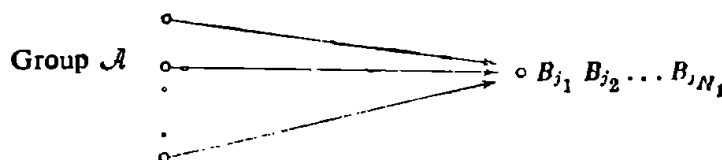


Fig. 30.

The group \mathcal{A} (consisting of $2^{H_{\alpha}(\mathcal{B})N_1}$ chains of transmitted signals corresponding to the fixed chain $B_{j_1}B_{j_2} \dots B_{j_{N_1}}$ received at the channel output) plays a central role in the method we shall use to decode the received message $B_{j_1}B_{j_2} \dots B_{j_{N_1}}$. If the indicated group \mathcal{A} contains only *one* of our code words, then we shall assume exactly this code word to have been transmitted. However, in the case in which \mathcal{A} contains *more than one* code word, or contains *no* code word, or finally the received N_1 -term chain does not belong at all to the collection of $2^{H(\infty)N_1}$ 'probable' chains $B_{j_1}B_{j_2} \dots B_{j_{N_1}}$, we shall assume any one code word chosen arbitrarily from the existing code words to have been transmitted (say, the code word with number 1 to have been transmitted in all these cases; it will be seen later that this specific agreement is in fact of no consequence).

Now we have already chosen the coding method (i.e., finding the 2^{HN} code words needed by us) and the decoding method (i.e., deciphering the received message). Hence, we can proceed to determine the *probability of decoding error*. Here, however, we are immediately confronted with one difficulty. Suppose the code word $A_{i_1}A_{i_2} \dots A_{i_{N_1}}$ to have been transmitted and the message $B_{j_1}B_{j_2} \dots B_{j_{N_1}}$ to have been received at the output. Let us now denote by P the probability that by using the decoding method described above we arrive at a wrong conclusion, i.e., conclude that some code word other than $A_{i_1}A_{i_2} \dots A_{i_{N_1}}$ was transmitted. It is clear that the quantity P , in principle, can be different for different code words $A_{i_1}A_{i_2} \dots A_{i_{N_1}}$; thus, for instance, our decoding method explicitly places the code word with number 1 in an exclusive setting. Is it necessary because of this that the quantity P be calculated separately for different code words (or separately only for the first and all remaining such code words)? We shall see below that the answer is negative, because we shall use estimates that remain valid for all code words without exception. But, besides, our decoding method depends also on the choice of code words to be used and this choice, as we know, is determined by the outcome of an experiment consisting of 2^{HN} draws of a paper from the urn, i.e., depends on certain random events. Hence P is also a *random variable* in the sense explained on p. 5. Such a

variable can have many different values; we shall calculate below just *the mean value* of P .

We know that if the number $N_1 = (H/c_1)N$ is sufficiently large, then the message $A_{i_1}A_{i_2} \dots A_{i_{N_1}}$ is transformed into one of the chains $B_{j_1}B_{j_2} \dots B_{j_{N_1}}$ of the group \mathcal{B} corresponding to this message with probability *arbitrarily close to unity*. Furthermore, we assume N_1 to be so large that the indicated probability is not *less than* $1 - (\epsilon/4)$, where ϵ is a preassigned small number. Suppose now that $B_{j_1}B_{j_2} \dots B_{j_{N_1}}$ is the 'probable' chain of received signals which belongs to group \mathcal{B} corresponding to some code words $A_{i_1}A_{i_2} \dots A_{i_{N_1}}$. Denote by Q the probability that the chain referred to also belongs simultaneously to a group \mathcal{B} that corresponds to *at least one more code word* (i.e., the probability that the group \mathcal{A} corresponding to our chain $B_{j_1}B_{j_2} \dots B_{j_{N_1}}$ contains, in addition to $A_{i_1}A_{i_2} \dots A_{i_{N_1}}$, at least one more code word). It is clear that both Q and P are random variables. Further, it is obvious that the received message $B_{j_1}B_{j_2} \dots B_{j_{N_1}}$ will certainly be decoded correctly if the following two conditions are satisfied :

- (A) This message belongs to the group \mathcal{B} corresponding to the transmitted code word.
- (B) Except for the above-mentioned group it does not belong to any of the groups \mathcal{B} corresponding to other code words used.

Hence an erroneous decoding can take place only if either the condition (A) or (B) is not satisfied. But we know that the probability of the sum $\bar{A} + \bar{B}$ of two events \bar{A} and \bar{B} (meaning respectively that the events A and B *do not take place*) does not exceed the sum of the probabilities of \bar{A} and \bar{B} (see pp. 9–10). Consequently, the probability of erroneously decoding the received N_1 -term chain must satisfy the inequality

$$P \leq \frac{\epsilon}{4} + Q;$$

here $\epsilon/4$ is greater than or equal to the probability that condition (A) is not satisfied (i.e., the event \bar{A} takes place) and Q is equal to the probability of (B) not being fulfilled (i.e., the probability of \bar{B}). In this inequality $\epsilon/4$ is a fixed number, but P and Q are random variables; for estimating the mean value of P it is, therefore, necessary only to estimate the mean value of Q .

Besides the code words $A_{i_1}A_{i_2} \dots A_{i_{N_1}}$, there exist also $2^{HN} - 1$ other code words. We renumber afresh these $2^{HN} - 1$ words in an arbitrary order and denote by a_i the random event such that the group \mathcal{A} corresponding to the chains $B_{j_1}B_{j_2} \dots B_{j_{N_1}}$ contains the i th *code word*. Condition (B) will not be satisfied iff at least one of the events $a_1, a_2, \dots, a_{2^{HN}-1}$ occurs; in other words, event \bar{B} equals the sum of events $a_1 + a_2 + \dots + a_{2^{HN}-1}$. But the probabi-

lity of the sum of events cannot exceed the sum of the probabilities of these events (see pp. 9-10), hence

$$Q \leq q_1 + q_2 + \dots + q_{2^{HN}-1},$$

where q_i is the probability of a_i .

Let us now try to determine the mean value of probability q_i . Since the i th code word is chosen at random (as are all the remaining code words), hence with the same probability $2^{-H(\beta)N_1}$ it can coincide with each of the $2^{H(\beta)N_1}$ existing 'probable' chains of N_1 transmitted signals A_i . In those $2^{H(\beta)N_1}$ cases in which the i th code word is found to coincide with one of the $2^{H(\beta)N_1}$ chains belonging to the group \mathcal{A} that corresponds to the chains $B_{j_1}B_{j_2} \dots B_{j_{N_1}}$, the event a_i takes place, i.e., its probability is unity; in the remaining $2^{H(\beta)N_1} - 2^{H(\alpha(\beta)N_1)}$ cases this event does not take place, i.e., its probability is zero. Thus, $q_i = 1$ for $2^{H(\beta)N_1}$ equally likely outcomes of an experiment consisting of the draw of $2^{H(\beta)N_1}$ papers from the urn and $q_i = 0$ for $2^{H(\beta)N_1} - 2^{H(\alpha(\beta)N_1)}$ the remaining equally likely outcomes. Hence, it is clear that

m.v. q_i

$$= \frac{2^{H(\alpha(\beta)N_1)}}{2^{H(\beta)N_1}} \times 1 + \frac{2^{H(\beta)N_1} - 2^{H(\alpha(\beta)N_1)}}{2^{H(\beta)N_1}} \times 0 = \frac{2^{H(\alpha(\beta)N_1)}}{2^{H(\beta)N_1}} = 2^{[H(\alpha(\beta)) - H(\beta)]N_1}.$$

But the mean value of all variables q_i is the same (because all numbers i are equivalent), and Q does not exceed the sum $2^{HN} - 1$ of the variables q_i ; hence the mean value of Q is not greater than

$$(2^{HN} - 1) \times 2^{[H(\alpha(\beta)) - H(\beta)]N_1} < 2^{HN} \times 2^{-\frac{H(\beta) - H(\alpha(\beta))}{c_1} HN} = 2^{-\left(\frac{c}{c_1} - 1\right) HN}.$$

We now recall that $c_1 < c$. This implies that for large N the expression appearing on the right-hand side of the preceding inequality is represented by the number 2 raised to a *negative power extremely large in absolute magnitude*, i.e., it is *quite small*. In particular, no matter how small the chosen number ϵ , N can be taken so large that this expression (and hence also the mean value of Q) is less than $\epsilon/4$.

But we know that $P \leq (\epsilon/4) + Q$; hence

$$\text{m.v. } P \leq \text{m.v. } Q + \frac{\epsilon}{4}.$$

But

$$\text{m.v. } Q < \frac{\epsilon}{4}$$

for sufficiently large N . Hence, choosing N sufficiently large, it can always be assured that the mean value of the probability P of erroneous decoding of any

of the 2^{HN} code words (which correspond to 2^{HN} 'probable' N -letter messages) is less than $\epsilon/2$, where ϵ is any preassigned (no matter how small!) positive number.

The result obtained facilitates the proof of Shannon's fundamental noisy coding theorem. For this we make use of the fact that the *mean value of any random variable cannot be less than all its possible values* (see pp. 6-7). In application to our case this means that *among the $(2^{H(\beta)N_1})^{2^{HN}}$ different possible choices of our 2^{HN} code words* (i.e., among all different outcomes of the experiment that consists of 2^{HN} successive draws of papers from the urn containing $2^{H(\beta)N_1}$ papers) *there is at least one for which the value of P is found to be less than $\epsilon/2$.*

The last assertion is quite close to the one we desire to prove but it is still inadequate for our purpose. The point is that P is the probability that *some fixed* transmitted code words $A_{i_1}A_{i_2} \dots A_{i_{N_1}}$ will be decoded erroneously at the channel output. It is, however, required to show that there exists some choice of the code words for which the probability of a decoding error when *any* of them is transmitted through a communication channel is less than ϵ . Denote now by P_i the probability of erroneous decoding of the transmitted i th code word. Then, $P_1, P_2, \dots, P_{2^{HN}}$ are the random variables, and mean value of each of them can be estimated in exactly the same way as the mean value of a variable fixed in them (denoted by P in the above discussion). Hence, the mean values of *all* variables P_i is less than $\epsilon/2$; but this still does not imply that for at least one of the choices at random of 2^{HN} code words the values of all variables $P_1, P_2, \dots, P_{2^{HN}}$ will be *simultaneously* less than $\epsilon/2$.

The preceding difficulty can, however, be circumvented by the following ingenious method. We choose at random not 2^{HN} chains $A_{i_1}A_{i_2} \dots A_{i_{N_1}}$, but *two times* their number, i.e., 2×2^{HN} chains. We take these 2×2^{HN} chains $A_{i_1}A_{i_2} \dots A_{i_{N_1}}$ as 2×2^{HN} code words and transmit them all through our communication channel, deciphering the received message $B_{j_1}B_{j_2} \dots B_{j_{N_1}}$ in exactly the same way as described above. Since $2 \times 2^{HN} = 2^{HN+1} = 2^{H_1N}$, where $H_1 = H + (1/N)$ for sufficiently large N differs from H as little as desired, it is easy to see that all preceding estimates also remain valid in this case. In other words, here also it can be shown that the mean value of the probability P of erroneous deciphering of the chain $B_{j_1}B_{j_2} \dots B_{j_{N_1}}$ received at the channel output, over which some one of our $2 \times 2^{HN} = 2^{H_1N}$ code words was transmitted, is for sufficiently large N necessarily less than $\epsilon/2$. Thus, if $P_1, P_2, \dots, \dots, P_{2 \times 2^{HN}}$ are the probabilities of erroneous deciphering of the 1st, 2nd, $\dots, \dots, 2 \times 2^{HN}$ th code words transmitted through the communication channel, then for sufficiently large N the mean values of all these variables are less than $\epsilon/2$.

We now consider a new random variable, namely

$$P_0 = \frac{P_1 + P_2 + \dots + P_{2 \times 2^{HN}}}{2 \times 2^{HN}},$$

equal to the *arithmetic mean* of all P_i . It is clear that if the mean values of all P_i are less than $\epsilon/2$, then the mean value of P_0 is also less than $\epsilon/2$. We now apply to the variable P_0 the assertion that the *mean value of a random variable cannot be less than all its values*. Then we obtain that, for at least one of the possible random choices of 2×2^{HN} code words, the value of P_0 must be less than $\epsilon/2$. However, all the variables $P_1, P_2, \dots, P_{2 \times 2^{HN}}$, which are the probabilities, cannot be negative; hence it is clear that if 2^{HN} or more of them are found to be not less than ϵ , then their arithmetic mean P_0 would not be less than $\epsilon/2$. This implies that *no less than 2^{HN} of the values of $P_i, i = 1, 2, \dots, 2 \times 2^{HN}$ must be less than ϵ* . The chains $A_{i_1}A_{i_2} \dots A_{i_{N_1}}$ corresponding to 2^{HN} suitable i (such that $P_i < \epsilon$) we take as 2^{HN} code words needed by us. In other words, we shall transmit them alone through our communication channel and decipher the received chains $B_{j_1}B_{j_2} \dots B_{j_{N_1}}$ as if no other code words existed. We now note that in all those cases in which for the received chains conditions (A) and (B) are found to hold in relation to 2×2^{HN} code words, they also remain all the more valid when half of the code words used previously are discarded. Hence all the inequalities derived above for the error probabilities P_i cannot worsen because of the fact that we rejected half of the initially chosen 2×2^{HN} code words. This gives the desired proof and establishes specially that *for sufficiently large N there always exists a choice of 2^{HN} code words $A_{i_1}A_{i_2} \dots A_{i_{N_1}}$ and of the method of decoding the received chain $B_{j_1}B_{j_2} \dots B_{j_{N_1}}$, such that the probability of decoding error is less than ϵ , irrespective of what code words were transmitted through the communication channel*.

The definition of the noisy channel capacity on p. 263 was based on the following assumption: if c is the largest amount of information that can be obtained at the channel output when one elementary signal is transmitted through the communication channel, then on receiving L such signals we cannot obtain more than Lc units of information. This assumption seems to be quite natural, but its mathematical proof is nevertheless not trivial. We shall now briefly explain how such a proof can be deduced.

Suppose that β (resp. α) is an experiment consisting of determining the value of one elementary signal transmitted through the channel (resp. received at the channel output). Then by assumption $I(\alpha, \beta) \leq c$. It is required to show that if $\beta_1\beta_2 \dots \beta_L$ is a compound experiment consisting of the successive realizations of experiments $\beta_1, \beta_2, \dots, \beta_L$ (i.e., consisting of the successive transmission of L elementary signals), and $\alpha_1\alpha_2 \dots \alpha_L$ is another compound experiment consisting of receiving these L transmitted signals, then necessarily

$$I(\alpha_1\alpha_2 \dots \alpha_L, \beta_1\beta_2 \dots \beta_L) \leq Lc.$$

It is clear that we only need to demonstrate that

$$I(\alpha_1\alpha_2 \dots \alpha_L, \beta_1\beta_2 \dots \beta_L) \leq I(\alpha_1, \beta_1) + I(\alpha_2, \beta_2) + \dots + I(\alpha_L, \beta_L).$$

In fact, each term on the right-hand side of this inequality equals the information about one transmitted signal contained in the corresponding received signal, i.e., it cannot exceed c .

We can restrict ourselves to the simplest case $L = 2$. This is possible since in the inequality obtained α_1 and β_2 can always be replaced by the compound experiments $\alpha_2\alpha_3 \dots \alpha_L$ and $\beta_2\beta_3 \dots \beta_L$ and then induction performed over the number L . As regards the proof of our inequality for $L = 2$, it can be obtained quickly by applying the *triple information equation* (see p. 92), which states that

$$I(\beta\gamma, \alpha) + I(\beta, \gamma) = I(\alpha\gamma, \beta) + I(\alpha, \gamma).$$

Putting $\beta = \alpha_1$, $\gamma = \alpha_2$ and $\alpha = \beta_1\beta_2$ in this equation we get

$$I(\alpha_1\alpha_2, \beta_1\beta_2) + I(\alpha_1, \alpha_2) = I(\beta_1\beta_2\alpha_2, \alpha_1) + I(\beta_1\beta_2, \alpha_2).$$

We now make use of the fact that the information contained in a compound experiment $\beta\gamma$ relative to a certain experiment α is equal to $I(\beta, \alpha)$ if the conditional probability of the outcome of α for a given outcome of $\beta\gamma$ in fact depends only upon the outcome of β (see p. 89). In our case the conditional probability of the outcome of α_1 , given the outcome of $\beta_1\beta_2\alpha_2$, can obviously depend only on the outcome of β_1 ; exactly in the same way, the conditional probability of the outcome of α_2 , given the outcome of $\beta_1\beta_2$, depends only on the outcome of β_2 . Hence

$$I(\beta_1\beta_2\alpha_2, \alpha_1) = I(\beta_1, \alpha_1), \quad I(\beta_1\beta_2, \alpha_2) = I(\beta_2, \alpha_2), \quad (C)$$

and since $I(\alpha_1, \alpha_2) \geq 0$ (the information is always nonnegative), we have

$$I(\alpha_1\alpha_2, \beta_1\beta_2) \leq I(\beta_1, \alpha_1) + I(\beta_2, \alpha_2),$$

giving the desired proof.†

We shall now show one more method of proving *Shannon's fundamental noisy coding*

†In deriving the equations (C) we have factually made use of the following result: the conditional probability of the outcome $B_k B_l$ of experiment $\alpha_1\alpha_2$ given that the experiment $\beta_1\beta_2$ has outcomes $A_i A_j$ (i.e., the probability of receiving a pair of signals $B_k B_l$ if $A_i A_j$ are transmitted) can be represented in the form

$$P_{A_i A_j}(B_k B_l) = P_{A_i}(B_k) \times P_{A_j}(B_l),$$

where $P_{A_i}(B_k)$ and $P_{A_j}(B_l)$ are the characteristics of the noisy channel, which are known to us. Indeed, this expressly implies that the outcome of α_1 (resp. α_2) depends only on the outcome of β_1 (resp. β_2). If we now substitute these probabilities $P_{A_i A_j}(B_k B_l)$ in the expression for the conditional entropy $H_{\beta_1\beta_2}(\alpha_1\alpha_2)$, then by elementary transformations it can be shown directly that

$$H_{\beta_1\beta_2}(\alpha_1\alpha_2) = H_{\beta_1}(\alpha_1) + H_{\beta_2}(\alpha_2),$$

and, consequently,

$$I(\alpha_1\alpha_2, \beta_1\beta_2) = H(\alpha_1\alpha_2) - H_{\beta_1\beta_2}(\alpha_1\alpha_2) \leq I(\alpha_1, \beta_1) + I(\alpha_2, \beta_2)$$

(since $H(\alpha_1\alpha_2) \leq H(\alpha_1) + H(\alpha_2)$; see p. 64). However, such a proof is found to be lengthier than the one ingeniously derived above.

theorem for the simplest binary symmetric channel.† Through such a channel we can transmit two elementary signals A_1 and A_2 , each of them having the probability $1 - p$ (resp. p) of receiving the same (resp. the opposite) signal at the output. As noted on p. 265, without restricting generality it can be assumed that $p < \frac{1}{2}$. Sequences $A_{i_1}A_{i_2} \dots A_{i_{N_1}}$ of N_1 signals are used as code words. Here all i_k (where $k = 1, 2, \dots, N_1$) can take the value 1 or 2, and hence there exist altogether 2^{N_1} such different sequences. Suppose that ϵ is some preassigned small number; the requirement is that the *probability of error in deciphering any transmitted code word does not exceed ϵ* . We are interested in how many code words can be chosen without coming into conflict with the italicized condition. It is shown below that for sufficiently large N_1 the possible number K of such code words can be made arbitrarily close to 2^{cN_1} , where

$$c = 1 + (1 - p) \log (1 - p) + p \log p$$

is the capacity of the channel under consideration related to one transmitted signal. Since a message that one fixed word is chosen by us from K possible words can supply $\log K$ bits of information, this implies that over the channel we can transmit information at a rate as close as desired to $C = Lc$ bits per unit time in such a way that the probability of error in deciphering each transmitted signal does not exceed ϵ . Therefore, the proof of the formulated statement is equivalent to the proof of Shannon's theorem.

For proof, the foremost requirement is to indicate a method of decoding the obtained collection of signals which ensures that the probability of error in deciphering each code word will not exceed ϵ . For this purpose, it is appropriate to make use of the Chebyshev inequality proved in Chap. 1.4. Using formula (****) on p. 35, it is easy to show that if

$$N_2 = \sqrt{2N_1p(1-p)/\epsilon},$$

then the probability p_0 that the error number x in decoding N_1 successively transmitted elementary signals A_i does not exceed $M = N_1p + N_2$ satisfies the inequality

$$p_0 = P(x < N_1p + N_2) > 1 - \frac{\epsilon}{2}. \quad (*)$$

We further note that for fixed p and ϵ the ratio

$$\frac{N_2}{N_1} = \sqrt{\frac{2p(1-p)}{\epsilon}} \times \frac{1}{\sqrt{N_1}}$$

can be made as small as desired if only N_1 is chosen sufficiently large. Hence

$$M = N_1p + N_2 = N_1(p + N_2/N_1)$$

can be made as close as desired to N_1p . In particular, when $p < \frac{1}{2}$ and N_1 is sufficiently large, $M = N_1p + N_2$ is less than $N_1/2$; hereafter N_1 will be taken to be so large that the preceding condition is satisfied.

We now choose the first code word (which for brevity we denote by A_1) in an arbitrary

†As already remarked above the idea of this proof is due to Feinstein who, however, studied directly the general case of an arbitrary communication channel. The application of Feinstein's arguments to the simplest particular case of a binary symmetric channel was examined by Gilbert [184] and Slepian [187]; one more variant of the simplified proof of Shannon's theorem for this case can be found in Barnard [180].

manner from among 2^{N_1} different chains $A_{i_1}A_{i_2}\dots A_{i_{N_1}}$. We shall consider A_1 to have been transmitted if at the channel output a message is received, differing from N_1 -term chain A_1 in not more than M elementary signals. We denote by the symbol $R(A_1)$ a collection of all possible N_1 -term chains differing from the chain A_1 in not more than M signals. Thus, the received N_1 -term chain is deciphered as the chain A_1 if it belongs to the collection $R(A_1)$; the probability of error in decoding A_1 then does not exceed $\epsilon/2$ because of (*).

We now take up the choice of the second code word A_2 . We first agree to regard A_2 to have been transmitted if at the channel output there is received an N_1 -term chain that

- (a) differs from A_2 in not more than M elementary signals; and
- (b) does not belong to the collection $R(A_1)$.

We are interested in only such code words A_2 , the probability of whose erroneous decoding at the channel output does not exceed ϵ . It is clear that this is certainly the situation *if in the transmission of the chain A_2 the probability of receiving some of the chains of the collection $R(A_1)$ is less than $\epsilon/2$* . For those cases in which an N_1 -term chain satisfying this condition does not exist at all, we consider that $K = 1$; if however, there exist N_1 -term chains satisfying it, we accept any of them as A_2 .

We act similarly also in the choice of the third code word A_3 . Namely, if there does not exist an N_1 -term chain of signals to be transmitted such that *when it is transmitted the probability of receiving at the channel output in place of it one of the chains belonging to either the collection $R(A_1)$ or $R(A_2)$ is less than $\epsilon/2$* , then we consider that $K = 2$; otherwise, we take any of the chains satisfying the italicized condition as the third code word A_3 . In analogy to this, after the first k code words A_1, A_2, \dots, A_k are chosen, as the $(k + 1)$ th code word we choose any N_1 -term chain A_{k+1} such that *in the case of its transmission through the communication channel the probability of receiving at the channel output one of the chains belonging to $R(A_1)$, or $R(A_2)$, \dots , or $R(A_k)$ is less than $\epsilon/2$* . The choice of all code words is regarded to be complete when it is found that no new chain satisfying the condition set out here exists. When decoding the messages received at the channel output, we regard the i th word A_i to have been transmitted if the received chain is the one that

- (a') differs from A_i in not more than M signals; and
- (b') belongs to none of the collections $R(A_1), R(A_2), \dots, R(A_{i-1})$.

If, however, the received chain differs from *all* existing code words A_1, A_2, \dots, A_K in more than M signals, then we decode it arbitrarily (say, we agree in all such cases to consider the code word A_1 to have been transmitted). It is clear that the method employed for decoding the received N_1 -term chains of signals guarantees that when any of the words A_1, A_2, \dots, A_K is transmitted it is decoded correctly at the channel output with probability exceeding $1 - \epsilon$. Thus, what remains to verify is just that the number K of such words for sufficiently large N_1 is sufficiently large (and, expressly, that K can be made as close as desired to $2^{\epsilon N_1}$).

In order to evaluate K , we first estimate the number L_0 of chains occurring in the collection $R(A)$ (where A is an arbitrary N_1 -term chain). It is clear that $R(A)$ includes

- (0) one chain A ;
- (1) $\binom{N_1}{1} = N_1$ distinct chains differing from A in one signal;
- (2) $\binom{N_1}{2}$ distinct chains differing from A in two signals;
- \dots
- (M) $\binom{N_1}{M}$ distinct chains differing from A in $M = N_1 p + N_2$ signals.

Hence

$$L_0 = 1 + \binom{N_1}{1} + \binom{N_1}{2} + \dots + \binom{N_1}{M}.$$

The number of terms on the right-hand side of this equality we estimate as the number $M = N_1 p + N_2 < N_1/2$ (because the term 1 at the start cannot affect an estimate for very large L_0). Moreover, it is known that in the sequence of binomial coefficients

$$\binom{N_1}{0} = 1, \binom{N_1}{1}, \binom{N_1}{2}, \binom{N_1}{3}, \dots, \binom{N_1}{N_1-1}, \binom{N_1}{N_1} = 1,$$

the terms monotonically increase up to the middle of this sequence. Hence, since $M < N_1/2$, the leading coefficient in the sequence $\binom{N_1}{1}, \dots, \binom{N_1}{M}$ is the *last* coefficient. Hence

$$L_0 < M \times \binom{N_1}{M} < \frac{N_1}{2} \times \binom{N_1}{M}.$$

Futhermore, using inequality (**) on p. 165 and noting that

$$N_1 - M = N_1(1 - p) - N_2 = N_1 q - N_2,$$

where $q = 1 - p$, we get

$$\begin{aligned} L_0 &< \frac{N_1}{2} \frac{N_1^{N_1}}{(N_1 p + N_2)^{N_1 p + N_2} (N_1 q - N_2)^{N_1 q - N_2}} \\ &= \frac{N_1}{2} \frac{1}{\left(p + \frac{N_2}{N_1}\right)^{N_1 p + N_2} \left(q - \frac{N_2}{N_1}\right)^{N_1 q - N_2}}. \end{aligned} \quad (**)$$

It is further required to estimate the number L_1 of all possible N_1 -term sequences of received signals occurring in *at least one* of the collections $R(A_1), R(A_2), \dots, R(A_K)$. We set forth our reasonings as follows. Let us consider the process of the transmission of all 2^{N_1} possible N_1 -term sequences $A_1, A_2, \dots, A_{2^{N_1}}$, each of these sequences having the same probability $1/2^{N_1}$ of being transmitted.† In such a case the probability of transmission of a sequence belonging to at least one of the collections $R(A_1), R(A_2), \dots, R(A_K)$ is obviously equal to $L_1/2^{N_1}$ (see

†The examination of such a transmission process occupies in the present proof a place allied to the role of the random coding procedure in Shannon's proof (see p. 292). Recall that for a binary symmetric channel the channel capacity is attained for the probabilities

$$p^0(A_1) = p^0(A_2) = \frac{1}{2}.$$

Hence the successive transmission of signals A_i , when any A_i is independent of all preceding signals and takes its possible values with probabilities p^0 , corresponds precisely to the transmission of all N_1 -term chains having the same probability $1/2^{N_1}$,

$R(A_2), \dots, R(A_K)$ turns out to be smaller than $\epsilon/2$, then in such a case this chain can be chosen as one more code word, contradicting the assumption that it is not possible to choose more than K code words.

Thus, on the right-hand side of the preceding multiline equality there occur 2^{N_1} columns, the sum of the terms of each of which is not less than $(1/2^{N_1}) \times (\epsilon/2)$. Hence, finally,

$$\frac{L_1}{2^{N_1}} \geq 2^{N_1} \left(\frac{1}{2^{N_1}} \times \frac{\epsilon}{2} \right) = \frac{\epsilon}{2}, \quad \text{i.e., } L_1 \geq \frac{\epsilon}{2} 2^{N_1}. \quad (***)$$

It is now quite straightforward to obtain the result we desire to prove. In fact, L_1 is the number of chains that belong to K different (in general, not disjoint) collections $R(A_1), R(A_2), \dots, R(A_K)$, each of which contains L_0 different chains. Consequently,

$$K \geq \frac{L_1}{L_0}.$$

Using the estimates (**) and (***) of L_0 and L_1 , it is found that

$$K > \frac{\epsilon}{N_1} 2^{N_1} \left(p + \frac{N_2}{N_1} \right)^{N_1(p + (N_2/N_1))} \left(q - \frac{N_2}{N_1} \right)^{N_1(q - (N_2/N_1))}.$$

We know that for sufficiently large N_1 the ratio N_2/N_1 becomes arbitrarily small. Since

$$\begin{aligned} \frac{\log K}{N_1} &> 1 + \left(p + \frac{N_2}{N_1} \right) \log \left(p + \frac{N_2}{N_1} \right) \\ &\quad + \left(q - \frac{N_2}{N_1} \right) \log \left(q - \frac{N_2}{N_1} \right) - \frac{\log N_1}{N_1} + \frac{\log \epsilon}{N_1}, \end{aligned}$$

it follows that $\log K/N_1$ for sufficiently large N_1 is larger than a number arbitrarily close to $c = 1 + p \log p + q \log q$. But we also know that the number K cannot be larger than 2^{cN_1} (see pp. 273 and 283); hence it is seen that for sufficiently large N_1 the number $\log K/N_1$ can be made as close as desired to c . As already remarked in the foregoing, it directly implies the validity of Shannon's theorem for a binary symmetric channel.

In conclusion we also present a rigorous proof of *Fano's inequality* (A') given on p. 287: in fact, the reasoning adduced on pp. 287-288 partially relies on intuitive notions about information and hence, strictly speaking, cannot be considered as a proof. Such a proof is easy to obtain if we attach exact meanings to all the arguments used earlier. We had based our arguments on the fact that the *amount of uncertainty of an experiment* β with n outcomes A_1, A_2, \dots, A_n having the probabilities $\pi_1, \pi_2, \dots, \pi_n$ is equal to the amount of uncertainty of an experiment γ consisting of verifying whether β had or did not have the outcome A_n , plus the product of $\pi_1 + \pi_2 + \dots + \pi_{n-1} = 1 - \pi_n$ and the amount of uncertainty of the experiment γ_1 with $n-1$ outcomes, which represents the same experiment β but with the auxiliary restriction that the outcome A_n had not taken place. However, if as usual we denote by

$$H(\pi_1, \pi_2, \dots, \pi_n)$$

the quantity

$$-\pi_1 \log \pi_1 - \pi_2 \log \pi_2 - \dots - \pi_n \log \pi_n,$$

there certainly exists a method for the choice of code words (i.e. specific 'blocks' formed of lengthy sequences of elementary signals) that allows information transmission at a rate v_1 such that the probability of erroneous decoding of any letter of the transmitted message is less than an arbitrary (but preassigned) number ϵ . On pp. 276–277 it was also remarked that Shannon's theorem can be formulated differently as follows: *if $c_1 < c$ and N is large enough, then $2^{c_1 N}$ code words of length N for sufficiently large N can be chosen in such a way that the probability of erroneous decoding of a sequence of N elementary signals received at the channel output is less than an arbitrary (preassigned) number ϵ regardless of what code word was actually transmitted.*[†] The latter version of the fundamental theorem is more apt in that it is related only to the *channel* but is in no way connected to the nature and statistical properties of the original message. Therefore, for the most part we shall use this version hereafter.

Shannon's coding theorem is fascinatingly simple but it also suffers from a serious shortcoming from the practical viewpoint. In fact, it is a typical 'existence theorem' and does not contain any indication of how one should choose code words of some acceptable length N in order to assure a sufficiently small probability of error versus a given quite high (i.e., quite close to $v = L(c/H)$) rate of transmission. The problem of determining a practically convenient method for the choice of code words for different noisy channels forms the content of *coding theory*, which developed after the appearance of Shannon's basic work [21] into a vast (and greatly important for application) independent discipline. A large variety of different approaches and methods are being used here, often borrowed from branches of modern mathematics that are seemingly highly abstract and detached from practical inquiries.^{††} Several tens, if not hundreds, of textbooks and monographs (of which [190], [193], [204], [209]–[212], and [215] are only a few examples) as well as several thousands of papers are devoted to the exposition of this science. Coding theory is also expounded in especial

[†]In Section 4.4 the length of code words was usually denoted by N_1 , since N was used there for denoting the length of encoded 'blocks' of the original lettered message. However, in the present section the original message is generally not considered; hence it will be convenient here to consider N as the code-word length.

^{††}This fact is reflected in the title of the interesting popular article [208] by the well-known American mathematician, N. Levinson, viz. 'Coding theory: a Counter-example to G.H. Hardy's Conception of Applied Mathematics'. The fact is that the famous English mathematician G. H. Hardy in his book, *A Mathematician's Apology*, written in 1940 (and subsequently reprinted many times), divided mathematics into 'pure' mathematics, which is a source of great aesthetic delight due to its harmony, logical regularity and elegance but is useless in practical life, and 'applied' mathematics that is needed for practice but is tedious and rather trite. It is precisely some of the most typical (according to Hardy's opinion) branches of 'pure' mathematics, (say) number theory or the theory of Galois fields that were later assigned a central rôle in (indisputably applied) coding theory!

sections of many general textbooks on information theory, applied algebra and combinatorics (see, for example, [2], [8], [11], [25], [191], [201] and [206]) and numerous review papers (for instance, [187], [195], [197], [208] and [219]). In the present text it is obviously impossible to cover even briefly just the fundamentals of modern coding theory. However, some relatively simple conclusions related to this theory can nevertheless be examined.

A few clarifications are useful as a starting point for understanding just the posing of problems in the coding theory. It is often asserted that all existing proofs of Shannon's fundamental theorem are *ineffective*, i.e., even in principle they cannot be used to determine a method that allows us to choose the code words (and a method of appropriately decoding the received sequence of elementary signals) that assure the low value of error probability for a given sufficiently high transmission rate. Actually, however, such an assertion cannot be regarded as completely valid.

Indeed, recall, for example, the method of proving Shannon's theorem by using 'random coding' described on pp. 292–297. In the course of this proof, it was suggested to choose randomly 2^{cN} code words of length N (out of certain preassigned $2^{H(\beta)N}$ 'probable' sequences of length N) and then it was shown that in such a case there exists a decoding method for which the mean value of the probability of erroneous decoding is sufficiently small (smaller than $\epsilon/2$). We further took advantage of the fact that it is always the case that at least one of the values of a random variable does not exceed its mean value; for proof of the theorem this was quite adequate for us. But it is also possible to make much further headway in this direction; it is clear that if the mean value of a non-negative random variable is quite small, then not one but *almost all* its values must be comparatively small. The latter circumstance finds its mathematical expression in the Chebyshev inequality (**) proved on p. 33. According to this inequality for any nonnegative random variable α

$$P(\alpha > c) < \frac{a}{c}, \quad \text{where } a = \text{m.v. } \alpha.$$

Hence, if $a = \text{m.v. } \alpha$ is so small that Ma also still remains small, where M is some comparatively large number, then the value of α does not exceed a small quantity Ma with very great probability (greater than $1 - 1/M$). Proceeding from similar arguments, it can be shown that if we make use of random coding (and the decoding method described on p. 293), then for sufficiently large N the probability of a decoding error (and not only its value for some specific but unknown choice of the 2^{cN} code words) is with very high probability (i.e., 'almost surely') extremely small. This gives us a seemingly very simple method for the choice of code words, which practically always leads to a small probability

of error.† For this, it is only necessary to take N sufficiently large and then choose randomly $2^{c_1 N}$ code words of length N (by means of the urn experiment described on p. 292).

But how can this 'simple' method actually be used in practice? Obviously, for obtaining good results here it is usually necessary to prescribe that N be at least of the order of many tens or even hundreds. If we assume that $N = 100$ and $c_1 = 0.5$, then it is necessary for us to choose randomly $2^{50} \approx 10^{15}$ distinct sequences of 100 elementary signals and all of them must be memorized. However, this itself is the smaller part of the assignment, for incomparably greater difficulty is encountered in decoding the received sequences of elementary signals. By what was stated on p. 278 et seq., for such decoding we must examine all 2^{50} groups \mathcal{B} corresponding to our code words to ascertain to which of them the received sequence of signals belongs and to which of them it does not, which poses a problem beyond the capabilities of all existing (and even those likely to appear in the near future) computers.

It is thus seen that the basic dilemma in coding theory is mainly that in general it is impossible to indicate a coding method (i.e., the method of choice of $2^{c_1 N}$ code words of length N) and a decoding method (i.e., a method of suitably deciphering the received sequences of N signals) that assures a high transmission rate and at the same time a small probability of error. The most essential requirement here is that both the coding and, what is particularly difficult, the decoding must be made comparatively simple in practice. It is not easy to meet this specification. This persisting difficulty has precisely motivated a vast number of investigations devoted to the development of various practically acceptable methods of coding and decoding, which even if not *optimal* (i.e., the best of all possible) are nevertheless sufficiently *good* (i.e., allow us to achieve a relatively high transmission rate without a large probability of error).

For the sake of simplicity, we confine ourselves to only a *binary* channel, i.e., we consider a channel over which we can transmit only two elementary signals, (say) on—off current, and such that the same two signals are obtained at the channel output. Denote by the digits 0 and 1 the signals to be used; in such a case all code words are sequences of these digits, i.e., *the numbers of a binary system*. Code words of length N must be chosen here from the set containing all 2^N distinct N -valued binary numbers, the sequences $a_0 a_1 \dots a_{N-1}$, where all a_i , $i = 0, 1, \dots, N - 1$, take the value 0 or 1. The collection of all chosen code words is now called a *code*. If we accept *all* 2^N distinct N -valued numbers as code words, then the information transmission rate will be the highest

†The term 'practically always' means here that the chosen code can fail only in the highly improbable case with 'exceptionally bad luck'. But if N is sufficiently large, then this possibility can be ignored. Moreover, even in the case of such failure the situation can be saved: if we are convinced (by means of a transmission test) that the chosen code is bad, it is possible to simply discard it and choose all code words afresh by means of the same method.

(namely, L bits/unit time, or equivalently L/H letters/unit time), but then there is no opportunity to determine at the channel output whether the transmission errors have taken place and how many, and specifically what signals have been received in error. If, however, we restrict ourselves to a smaller number of code words, then the resultant 'code redundancy' can be used for further transmission of some information about the distortions induced in the channel. Thus, for instance, we can use to advantage the simplest method of N -multiple repetition of each elementary signal (i.e., employ as a code only the two simplest code words $00 \dots 0$ and $11 \dots 1$ of length N), and decode a sequence of length N received at the channel output as $00 \dots 0$ if it contains more 0's than 1's and as $11 \dots 1$ otherwise. It is clear that such a transmission method when N is sufficiently large (and subject to the conventional restriction that the probability of the distortion of an elementary signal in the process of its transmission is less than $\frac{1}{2}$) assures us of quite low probability of erroneous decoding of a transmitted message. However, the transmission speed will also be quite low here (during the time N/L required for the transmission of N elementary signals, only 1 bit of information is transmitted, which corresponds to a transmission rate of L/N bits/unit time $= L/HN$ letters/unit time). It is natural that in many cases we shall not be able to manage with such a low transmission speed. Hence, the classes of codes intermediate between the two extreme codes considered above are of greatest interest to us. Such intermediary codes are amenable to rather high transmission rate and simultaneously allow us to correct many distortions in the transmitted message.

The simplest method of increasing the transmission reliability by a multiple repetition of each elementary signal allows us to explain some important notions of coding theory. A code is called the *error-detecting code* if it permits to detect transmission errors, and the *error-correcting code* if it permits not only to detect an error, but also to determine this error, i.e., to reconstruct correctly the transmitted signal. It is clear that error-correcting codes are more useful than error-detecting codes, but usually the latter codes are much simpler. If, however, the probability of error is small, then even the possibility to detect the error is of great value. In fact, if it is known that errors are involved in the reception, then we can simply ignore the obtained message or, if acceptable, require the transmission to be repeated. It is clear that (say) triple repetition code, which codes elementary signals 0 and 1 as triplets 000 and 111 and decodes the received triplets according to the 'majority rule' (i.e., for example, 000 and 010 are decoded as 0 and 110 as 1), allows to correct any single error (but not double error) and to detect any single or double (but not triple) error. Let us suppose that this code is used for transmission through a binary symmetric channel sketched in Figure 18. If the error probability p is equal to 0.01 (i.e., 1% of the transmitted signals is received in error), then our method of coding makes the probability of error-free deciphering of each triplet as high as 0.9997 (i.e., the frequency of errors is close to 0.03%). The probability of detecting an error

becomes here close to 0.999999 (i.e., the frequency of a missed error is close to 0.0001%). Of course, the code triples the time of transmission, but nevertheless it is clear that it can be quite useful when error probability p is low and the time for transmission is not of great importance.

The error-detecting and error-correcting properties of code considered above are closely connected to the fact that the code uses only a small part of all three-term signal sequences as code words and the selected code words differ here considerably from each other. In general, it is clear that if all code words have the same length N and any two of them differ not less than in d elementary signals, then the code permits to detect in a block of N signals any number of errors which is less than d . In fact, such number of errors certainly changes the transmitted code words into a block of N elementary signals which is not a code word at all. Moreover, if we decode any received block of N signals as a code word differing from it in the least number of elementary signals (or one of such code words, if there are several of them), then we shall correct any number of errors which is less than $d/2$. (This is clear since two distinct code words, both of which differ from a received block in less than $d/2$ elementary signals, cannot exist.) In the above example of a code with two code words 000 and 111, evidently, $N = 3$ and $d = 3$; hence the code permits to detect any number of errors which is less than 3 (i.e., equal to 1 or 2) and to correct the errors whose number is less than $\frac{3}{2}$ (i.e., equal to 1).

The multiple repetition method for increasing the transmission reliability is in fact used seldom since it is quite far from being optimal. The following comparatively general method of the use of code word redundancy for the transmission of information about the distortions is used much more frequently. The number of code words of length N chosen here is 2^{N-1} (i.e., is equal to half of the number of all distinct sequences of N binary signals). Let us agree to form 2^{N-1} code words of all possible sequences $a_0 a_1 \dots a_{N-2}$ of $N - 1$ digits 0 and 1, but the N th digit a_{N-1} is so chosen every time that the sum $a_0 + a_1 + \dots + a_{N-1}$ is *even*. In such a case, the presence of a *single* error (i.e., when one of the received N elementary signals is in error) leads to the emergence of sequences $a' a' \dots a'_{N-1}$ at the output such that the sum $a'_0 + a'_1 + \dots + a'_{N-1}$ is odd (since the possible distortion is that either 0 is taken for 1, or 1 for 0). This position enables us to detect easily the presence of a single error, even though it does not allow us to ascertain what specific signal is received in error (precisely, the property of the sum $a'_0 + a'_1 + \dots + a'_{N-1}$ being odd indicates that an *odd* number of signals has certainly been received in error, but the code does not permit the even number of errors to be detected). Nevertheless in those cases for which in the transmission of N signals, the probability of the appearance of more than one error is extremely low, the highly simple coding method described here is indeed of great value. In fact, if it is known with certainty that errors are involved in the reception, then we can simply ignore the obtained message, or if we wish we may require the transmission to be repeated. On the

other hand, the transmission rate in such a coding method still remains quite high; with a maximal value of L bits/unit time it decreases altogether to just $[(N-1)/N]L$ bits/unit time $= [(N-1)/N] (L/H)$ letters/unit time.

The 'parity check' method described above can also be applied *several times*, and this enables us in many cases not only to detect the presence of an error but also to correct it. Consider, for instance, the case in which $N = 3$ and the number of code words to be employed is 2. It is clear that in such a case it is reasonable to choose the triples 000 and 111 as code words; such a choice from the viewpoint of using 'parity checks' can be justified as follows. We form two code words on the basis of two possible values of the first elementary signal a_0 (i.e., we consider that only the signal a_0 actually contains the information). Furthermore, we agree to transmit after each 'information' signal a_0 two more 'check' signals a_1 and a_2 so chosen that both the sums $a_0 + a_1$ and $a_0 + a_2$ are even (it is easy to see that this precisely reduces to the choice of the frequencies 000 and 111 as code words). In such a case it is seen that if only in the received triple signals two or three errors do not occur (i.e., if only a correct transmission and a transmission with a single error are considered possible), then by the parity check of the sums $a'_0 + a'_1$ and $a'_0 + a'_2$ in the triplet $a'_0 a'_1 a'_2$ received at the output it can be ascertained without error what specific triplet was actually transmitted. In fact, if both the sums $a'_0 + a'_1$ and $a'_0 + a'_2$ are found to be even, then it directly implies that there is no transmission error (recall that the possibility of double error is excluded). If, however, only one of them is odd, then this means that the check signal a_1 or a_2 occurring in this sum is received in error, but if both the sums $a'_0 + a'_1$ and $a'_0 + a'_2$ are odd, then this implies that the information signal a_0 is received in error. Thus, at the price of decreasing the transmission rate by a factor of 3 (as compared to the maximal rate L bits/unit time) it is possible to achieve correction of all *single* errors in triplets of elementary signals.

The result derived above is obviously trivial (it is clear that by taking the triplets 000 and 111 as code words, we can achieve correction of all single errors), but it can be extended also to cases of many larger values of N . Thus, for instance, if $N = 7$ and the number of code words is $16 = 2^4$, then we can take the first four signals a_0, a_1, a_2 and a_3 as information signals (since the number of distinct quadruples $a_0 a_1 a_2 a_3$ is exactly sixteen), and choose the last three 'check signals' a_4, a_5 and a_6 such that the sums

$$s_1 = a_0 + a_1 + a_2 + a_4, \quad s_2 = a_0 + a_1 + a_3 + a_5,$$

and

$$s_3 = a_0 + a_2 + a_3 + a_6$$

are even. Here the 'parity check' of the three sums s_1, s_2 and s_3 at the channel output also allows us to determine uniquely whether an error has been admitted

in receiving (subject to the condition that the possibility of two or more errors in receiving seven signals is ignored) and, if it has, then in which signal it is included. In fact, if one of the 7 signals is received in error, then at least one of the sums must surely be found to be odd, so that the parity of the three sums positively indicates that there has been no transmission error. Furthermore, only one sum will be odd in that (and only that) case in which one of three 'check signals' (a_4 , a_5 or a_6) occurring in the sum is received in error. Finally, the non-parity of two of the three sums s_1 , s_2 and s_3 means that out of a_1 , a_2 and a_3 that signal which occurs in both these odd sums is received in error, and the non-parity of all the three sums implies that the first signal a_0 , occurring in all the sums, is received in error. It is easy to see that the 16 code words of length 7 in the given case have the form

0000000,	1000111,	0100110,	1100001,
0010101,	1010100,	0110011,	1110100,
0001011,	1001100,	0101101,	1101010,
0011110,	1011001,	0111000,	1111111.

The use of these code words yields the transmission rate

$$\frac{4L}{7} \text{ bits/unit time} = \frac{4L}{7H} \text{ letters/unit time,}$$

and at the same time allows us to correct all *single* errors (but not errors of higher multiplicity!) in 'blocks' of six elementary signals.

The corresponding code is, of course, not the 'best possible' but since both coding and decoding are carried out here without much difficulty, it fully justifies its practical usefulness. Let us consider again, for instance, a binary symmetric channel, in which the probability of receiving in error each of the two employed elementary signals is 0.01. The capacity of such a channel is given by

$$C = 0.92L \text{ bits/unit time}$$

(see p. 265). Hence, here definitely exists a code that allows us to transmit $0.92L$ bits of information per unit time and is such that the probability of a decoding error is less than an arbitrary preassigned number ϵ (which can be chosen as small as desired). However, how to construct such a code we do not know; furthermore, if ϵ is taken extremely small, then apparently code words of corresponding code will be very lengthy and the code itself will be extremely complex. Let us now try to use the very simple code described above with $N = 7$, in which to every four signals to be transmitted are added three further

check signals. Here, we transmit information at the rate

$$\frac{4}{7} L \approx 0.57L \text{ bits/unit time,}$$

which is appreciably lower than the limiting rate of transmission without error; in addition, the probability of a coding error here is obviously not 'as small as desired' but is equal to the probability that out of seven transmitted elementary signals two or more are received in error. Starting from here, it can be calculated that in such a transmission method in a sequence of 'elementary information signals' slightly less than 0.001th of the signals are obtained in error at the channel output so that the probability of receiving one elementary signal in error is here slightly below 0.001. It is seen that the probability of receiving one elementary signal in error is reduced in this case to less than $\frac{1}{128}$ th that for transmission not using 'check signals'. Since in this case both coding and decoding are highly straightforward and can be easily automatized, from a practical viewpoint the use of the described code unquestionably merits consideration.

It may be noted further that the examples described here of 'single-error-correcting codes' are quite intimately related to the content of the problem analyzed on pp. 107-108, in which it was supposed that among n given numbers either one number or none was thought of and it was required by means of the least number of questions (answers to which could be only 'yes' or 'no') to clarify whether or not a number was thought of and if yes, what number specifically. It is now convenient that instead of n numbers we consider N indices $0, 1, \dots, N-1$ appearing as subscripts to the code word $a_0 a_1 \dots a_{n-1}$; such a substitution obviously does not affect our arguments. By what is stated in our exposition on p. 108, it is required here to put not less than $\log(N+1)$ and not more than $\log(N+1) + 1$ questions; but our 'parity checks' are in fact equivalent to some questions (since each check can give two results: 'even' or 'odd', in analogy to 'yes' or 'no' answers to a question). In Chap. 3 answers to the questions contain definite information about the number thought of, since these were put by a person to whom this number was known. Similarly, in order that the result of a 'parity check' contains information about the possible distortion in transmission, it is necessary to know in advance that the sum of the signals to be transmitted is even or odd. Since in general it cannot be known what signal is transmitted, the preceding condition can be satisfied if and only if each sum to be transmitted contains at least one 'check signal', which we agree beforehand to choose such that the corresponding sum is found to be (say) even. It is thus clear that the number of 'check signals' that must be added coincides with the minimal number of 'parity checks' needed, i.e., it is equal to the number of those questions of which we spoke on p. 108. If, for instance, $N = 3$, then the number of questions cannot be less than $\log(3+1) = \log 4 = 2$; this also corresponds exactly to the fact that in the example of the single error-cor-

recting code described on p. 310, each 'information signal' a_0 to be transmitted is adjoined to *two* additional 'check signals' a_1 and a_2 . We further note that since the signals a_1 and a_2 are so chosen that the sums $a_0 + a_1$ and $a_0 + a_2$ are even, the parity checks of the corresponding sums at the channel output are equivalent to the answers to the questions of 'whether or not the pair of received signals a_0 and a_1 contains an error' and of 'whether or not the pair of signals a_0 and a_2 contains an error'. It is clear that answers to such questions allow us to determine uniquely any single error. In analogy to this, if $N = 7$, then the number of required questions (i.e., 'parity checks' and 'check signals') cannot be less than $\log(7 + 1) = \log 8 = 3$; this is exactly what is shown on pp. 310–311. Taking recourse there to the parity check of sums s_1 , s_2 and s_3 is equivalent to the answers to the questions of 'whether or not the received signals a_0 , a_1 , a_2 and a_4 contain an error?', 'whether or not the signals a_0 , a_1 , a_3 and a_5 contain an error?' and 'whether or not the signals a_0 , a_2 , a_3 and a_6 contain an error?'. It is obvious that answers to these questions also uniquely determine the erroneous signal if it exists.

In the general case of code words of length N , the number K of 'check signals' of a code needed to correct all single errors, must satisfy, by what is stated above, the inequality

$$\log(N + 1) \leq K < \log(N + 1) + 1,$$

so that

$$2^{K-1} - 1 < N \leq 2^K - 1.$$

The number of 'information signals' here is then equal to $N - K$. A code that uses code words of length N which consist of $M = N - K$ 'information signals' and K 'check signals', carrying no information but used for parity checks, we call an (N, M) -code.[†] The information transmission rate associated with such code is obviously $L(M/N)$ bits/unit time $= L(1 - K/N)$ bits/unit time. In the case considered $K < \log(N + 1) + 1$, so that K for large N is considerably smaller than N ; hence the transmission rate for large N is here quite close to the maximal rate of L bits/unit time. We see that the code under consideration when N is large assures a quite high transmission rate. Obviously, it is nevertheless not preferable to choose an extremely large N , because in that case the probability of the presence of *several* (more than one) errors in a block of N signals is sharply increased, i.e., the reliability of the code is reduced. In practice we have to resort to a compromise and choose some intermediate (neither

[†]Therefore (say) the Shannon-Fano code, or the Huffman code is not an (N, M) -code but the triple repetition code described above is $(3, 1)$ -code. General (N, M) -codes are often called also *block codes*. It is clear that $N > M$ for all error-detecting and error-correcting block codes. However, the case $N = M$ is widely used in cryptography where the coding is used only to make the message unintelligible for the uninitiated.

exceedingly large, nor yet too small) value of N . Also, the method of choice of 'check signals' for a general (N, M) -code, where $M = N - K$, correcting all single errors, can be set up via the path indicated on p. 108 for the problem of guessing a thought of number; we shall not dwell upon this here, because we are going to indicate below an entirely different method for the construction of the required code. We may further remark that the case of a single-error-correcting $(7, 4)$ -code analyzed on pp. 310–311 was considered by Shannon [21] by way of an example; the general single-error-correcting (N, M) -codes were examined in 1950 by Hamming [203] and are usually called the *Hamming codes* in the literature.†

Similarly, we can also approach the problem of the construction of a *double-error-correcting code* which corrects all single and double errors. Assume (say) that $N = 5$, and that we ignore the possibility of the simultaneous distortion of more than two of the five signals but prescribe that the code enables us to correct all distortions for the case in which their number does not exceed two. This setup leads us to the problem of determining $n \leq 2$ thought of numbers among some five numbers. By what was stated on p. 108 for determining these numbers, it is required to pose not less than

$$\log \left[\binom{5}{2} + \binom{5}{1} + 1 \right] = \log (10 + 5 + 1) = \log 16 = 4$$

questions; hence this specifies that at least four parity checks must be carried out and implies that of every five signals a_0, a_1, a_2, a_3 and a_4 at least four must be 'check signals'. It is not difficult to see that in the given case four check signals indeed suffice for solving the problem. It is, for example, possible to choose these signals a_1, a_2, a_3 and a_4 with the restriction that the sums

$$s_1 = a_0 + a_1, \quad s_2 = a_0 + a_2, \quad s_3 = a_0 + a_3 \quad \text{and} \quad s_4 = a_0 + a_4$$

be even. In such a case the parity of all sums considered at the receiving end of the channel means the absence of errors; the nonparity of one sum s_i implies the only corresponding signal a_i to be in error; the nonparity of two sums s_i and s_j , the signals a_i and a_j to be in error; the nonparity of three sums, say all except s_i , the signals a_0 and a_i to be in error; the nonparity of all four sums, only the

†It is, however, quite common to refer to only such single-error-correcting (N, M) -codes, in which $N = 2^K - 1$, as the Hamming codes (i.e., those in which N takes the greatest value possible for a given number K of 'check signals'). These codes possess significant properties, which we shall state at the close of this section. It is interesting to note that such $(2^K - 1, 2^K - K - 1)$ -codes were examined as early as 1942 (i.e., prior to the appearance of Hamming's and even Shannon's works) by the famous English statistician R. A. Fisher (see Berlekamp [190], Section 1.3) in an entirely different context (not formally connected to coding theory but equivalent to it).

signal a_0 to be in error.†

In the general case of a *double-error-correcting* code with an arbitrary number N of signals in every code word, the results derived on p. 108 show in exactly the same way that *the number K of 'check signals' and the 'parity checks' corresponding to them must satisfy the inequality*

$$K \geq \log \left[\binom{N}{2} + \binom{N}{1} + 1 \right] = \log \frac{N^2 + N + 2}{2}. \quad (*)$$

However, the question as to which specific 'check signals' should be chosen here (i.e., which 'parity checks' take us fastest to our goal) is not easy to answer in this case and thus a solution of the corresponding problem of guessing a number does not yield a general method of effectively constructing a suitable 'error-correcting code'. In analogy to this, in a still more general case of codes that enable us to detect and correct in a sequence of signals of length N *any number of errors not exceeding a given n* , the reasonings deduced on p. 108 allow us to say that *the number K of 'check signals' (and the 'parity checks' corresponding to them) required for this sequence must satisfy the equality*

$$K = \log \left[\binom{N}{n} + \binom{N}{n-1} + \dots + 1 \right]. \quad (**)$$

This straightforward conclusion is due to Hamming [203], and hence inequality (**) for the number K is frequently called the *Hamming inequality* or the *Hamming lower bound* on the number of 'check signals' of an n -error-correcting code. If $n = 1$, then the Hamming inequality (**) reduces to the result $N \leq 2^K - 1$ already known to us; here equality is attained for the Hamming codes with $N = 2^K - 1$. But the arguments set forth on pp. 107–108 in the general case do not indicate how we should choose the 'parity checks' we need (i.e., how to construct a code with the requisite properties); furthermore, they do not even allow us to state that for any K satisfying the Hamming inequality (**) there indeed exists a 'parity-check code' that contains K check signals and enables us to correct any number of errors less than n in a 'block' of N signals (in fact, for certain K satisfying this inequality, it is impossible to construct the requisite code). An estimate of the number of K 'check signals' that is clearly *sufficient* for it to

†It is easy to comprehend that the 'parity checks' described are equivalent to the answers to questions: 'shall the number of errors be even—when the signals a_0 and a_1 are received?'; 'when the signals a_0 and a_2 are received?'; 'when the signals a_0 and a_3 are received?'; and finally, 'when the signals a_0 and a_4 are received?' Here the answer to the first question separates from 16 distinct possible 'outcomes' of the transmission, in which not more than two elementary signals are distorted, a group of 8 admissible outcomes, i.e., contains the largest possible information; in the same way, all succeeding questions also extract exactly half of the remaining number of these possible 'outcomes'.

be possible to detect and correct any number of errors less than n in a block of N signals obtained from completely variant arguments is due to Varshamov [218], who showed that *for*

$$K > \log \left[\binom{N-1}{2n-1} + \binom{N-1}{2n-2} + \dots + 1 \right] \quad (***)$$

we can always construct a 'parity check code' having the requisite properties. The inequality (***) (making sharper the preceding related result due to Gilbert [201]) is called the *Varshamov-Gilbert inequality* or the *Varshamov-Gilbert upper bound* on a number K of check signals of an n -error-correcting code; a simple proof of this will be given later in this section. If $n > 1$, the Varshamov-Gilbert upper bound is in general found to be greater than the Hamming lower bound. Thus, here, there are the values of the number K of 'check signals' for which the corresponding inequalities do not exclude the possibility that an n -error-correcting $(N, W - K)$ -code exists, but at the same time they do not allow us to assert that such a code necessarily exists. In addition, all proofs of the Varshamov-Gilbert inequality, although relying on a definite method of constructing the required code, make no claim to the effect that this method can be successfully applied in practice. In fact, the existing construction proofs are found to be completely inadmissible for actual use (they are all based on sorting out of an enormous number of possibilities).

Even for the simplest case in which $n = 2$, a practicable method of constructing a 'parity-check code' that allows the correction of any *single* or *double* error in a block of an arbitrary number N of signals was not found until nearly ten years after the appearance of Hamming's work [203] describing a general single-error-correcting code. In this connection, the reader is referred to Bose and Ray-Chaudhuri [192] and Hocquenghem [204] where, surprisingly, the tools used for this purpose are found to belong to a subtle and quite complicated mathematical apparatus involving abstract algebra. We shall revert to Bose-Chaudhuri-Hocquenghem codes at the end of this section. A subsequent generalization of these methods, allowing us to construct codes correcting *any number of errors less than a given number n* , proved to be comparatively simple and was obtained at practically the same time the codes correcting *not more than two errors* were determined.

In order to give an idea of the method of constructing codes correcting not only single but also double (or generally multiple, not exceeding a given multiplicity) errors by the parity-check results, the first prerequisite is to define rigorously the notion of a 'parity-check code'. From this objective, a convenient starting point is to regard all arithmetic operations with the numbers 0 and 1 as operations that can have only two possible results, 0 and 1 symbolizing the fact that as a result of the operations we obtain an *even* number and an *odd* number, respectively. This leads us to the accompanying table listing the results

of all possible arithmetic operations carried out on the numbers 0 and 1:

$$\begin{aligned} 0 + 0 &= 0, & 0 + 1 &= 1, & 1 + 0 &= 1, & 1 + 1 &= 0; \\ 0 \times 0 &= 0, & 0 \times 1 &= 0, & 1 \times 0 &= 0, & 1 \times 1 &= 1. \end{aligned}$$

It is easy to see that the operations so obtained are 'addition' and 'multiplication' (which we call *addition* and *multiplication in an arithmetic with two symbols*, or, in short, in *2-arithmetic*), satisfying all rules of ordinary arithmetic.[†] This fact manifests itself if we say that a collection of two numbers 0 and 1, for which the addition and multiplication conventions in 2-arithmetic are defined, forms a *field* of two elements (or a *binary field*; cf. Appendix II).^{††}

It is now easy to describe a general (N, M) -parity check code. It is characterized by $K = N - M$ relations of the form

$$\begin{aligned} a_M &= b_{M,0}a_0 + b_{M,1}a_1 + \dots + b_{M,M-1}a_{M-1}, \\ a_{M+1} &= b_{M+1,0}a_0 + b_{M+1,1}a_1 + \dots + b_{M+1,M-1}a_{M-1}, \\ &\cdot \qquad \qquad \qquad \cdot \qquad \qquad \cdot \\ &\cdot \qquad \qquad \qquad \cdot \qquad \qquad \cdot \\ a_{N-1} &= b_{N-1,0}a_0 + b_{N-1,1}a_1 + \dots + b_{N-1,M-1}a_{M-1}. \end{aligned} \tag{1}$$

Here all coefficients

$$b_{M,0}, b_{M,1}, \dots, b_{M,M-1}, \dots, b_{N-1,0}, \dots, b_{N-1,M-1}$$

are elements of our field of two elements (i.e., the number 0 or 1), and all arithmetic operations appearing in these equations are understood in the 2-arithmetic sense (so that each equation means only that its left-hand and right-hand sides

[†]Strictly speaking, 'multiplication' in 2-arithmetic can be written without quotation marks since it does not differ from customary multiplication. Contrarily, 'addition' in 2-arithmetic varies from ordinary addition because here $1 + 1 = 0$ (because of which this addition is sometimes denoted by a special symbol, e.g. $\dot{+}$ or $\hat{+}$).

^{††}The fact that a collection of distinct elementary signals can be considered as a collection of all possible elements of a certain finite *field* is of great importance for all of modern algebraic coding theory. However, in algebra it is demonstrated that a *field with a given number m of distinct elements exists if and only if m is a power of a prime number* (i.e., equals p^k , where p is prime; see Appendix II). Hence, algebraic coding theory can be applied directly to a non-binary linear channel (which we shall not consider here at all, however) only in the case in which the number m of distinct elementary signals that can be transmitted over the channel have the form p^k . If this is not so, then we have to take further recourse to some tricks (for instance, we might never use some of the admissible signals).

taken in the ordinary sense have the same parity). The K parity checks corresponding to our (N, M) -code are parity checks of the sums of check signals a_i (where i takes $K = N - M$ values $M, M + 1, \dots, M + K - 1 = N - 1$) and those of the information signals a_0, a_1, \dots, a_{M-1} which correspond to coefficients $b_{i,0}, b_{i,1}, \dots, b_{i,M-1}$ equal to 1 (but not 0!).† For defining a code, it suffices to indicate all coefficients $b_{i,j}$ entering the equations set forth. It is appropriate here that all the left-hand sides $a_M, a_{M+1}, \dots, a_{N-1}$ in these equations be transferred first to the right-hand sides (taking account of the footnote on this page), and then that all coefficients in the obtained equations be arranged in the form of a table of $K = N - M$ rows and N columns, at the intersection of whose i th row and j th column appears the coefficient of a_j in the i th of our equations. It is easy to see that such a table has the form

$$\begin{bmatrix} b_{M,0} & b_{M,1} & & b_{M,M-1} & 1 & 0 & \dots & 0 \\ b_{M+1,0} & b_{M+1,1} & \dots & b_{M+1,M-1} & 0 & 1 & \dots & 0 \\ & & & & & & & \cdot \\ & & & & & & & \cdot \\ & & & & & & & \cdot \\ & & & & & & & \cdot \\ b_{N-1,0} & b_{N-1,1} & \dots & b_{N-1,M-1} & 0 & 0 & \dots & 1 \end{bmatrix}. \quad (2)$$

A rectangular array of m rows and n columns is called in mathematics a *matrix* of m rows and n columns, or briefly an $(m \times n)$ -matrix; thus, a general (N, M) -parity-check code is given by a $(K \times N)$ -matrix of 0's and 1's of the specific form (2). A collection of all possible code words of such a general (N, M) -parity-check code can easily be described thus: the information signals a_0, a_1, \dots, a_{M-1} can be arbitrary here (i.e., each of them can take, regardless of the others, both the values 0 and 1), but the check signals $a_M, a_{M+1}, \dots, a_{N-1}$ are uniquely defined by the information signals with the aid of equations (1), understood in the sense of 2-arithmetic. The total number of distinct code words in this case is obviously $2^M = 2^{N-K}$.

Note that sometimes a *parity-check code* is also defined rather more broadly as a collection of N -term sequences $a_0 a_1, \dots, a_{N-1}$ of symbols 0 and 1 such that

†Recall that in 2-arithmetic $1 + 1 = 0$ and hence $-1 = 1$. Therefore, when a term is carried over here from one side of the equation to the other side it does not necessarily change its sign, and the equation $x = y$ can be written both as $x - y = 0$ and $x + y = 0$ (both the relations are equivalent to each other, implying just that x and y have the same parity).

the numbers a_0, a_1, \dots, a_{N-1} satisfy K relations of the form

$$\begin{aligned} b_{M,0}a_0 + b_{M,1}a_1 + \dots + b_{M,N-1}a_{N-1} &= 0, \\ b_{M+1,0}a_0 + b_{M+1,1}a_1 + \dots + b_{M+1,N-1}a_{N-1} &= 0, \\ &\vdots \\ b_{N-1,0}a_0 + b_{N-1,1}a_1 + \dots + b_{N-1,N-1}a_{N-1} &= 0 \end{aligned} \quad (1')$$

(where the coefficients again take only the values 0 and 1, and the equations are understood in the sense of 2-arithmetic). The matrix corresponding to the most general code (1') is an arbitrary $(K \times N)$ -matrix consisting of 0's and 1's. Bearing in mind this broader definition, a particular code given by relations of the form (1) and matrices of the form (2) is called a *systematic parity-check code*. It is not difficult to show, however, that an arbitrary parity-check code can always be written as a systematic code with the number of 'check signals' not exceeding the number K of relations (1') (see Appendix II). Hence, as a rule, in the following we shall speak only of systematic codes.

In the literature on coding theory, parity-check codes are also often called *linear codes* or *group codes*. Both these terms are related to auxiliary properties of the considered codes, which are of interest in their own right and highly important if it is desired to carry over the theory of such codes to more general nonbinary channels (for which the notion of parity check has obviously no direct meaning). In order to clarify what these properties consist of, it is necessary to consider the operations of addition and multiplication by a number z (belonging to our field of two elements, i.e., equal to either 0 or 1) of N -tuple blocks $\mathbf{a} = (a_0, a_1, \dots, a_{N-1})$ of 0's and 1's. These operations can be conventionally described as follows:

$$\begin{aligned} (a_0, a_1, \dots, a_{N-1}) + (a'_0, a'_1, \dots, a'_{N-1}) &= (a_0 + a'_0, a_1 + a'_1, \dots, a_{N-1} + a'_{N-1}), \\ z \times (a_0, a_1, \dots, a_{N-1}) &= (za_0, za_1, \dots, za_{N-1}). \end{aligned}$$

We note incidentally that since all arithmetic operations are understood here in the sense of 2-arithmetic, the operation of multiplication of a block by a number is of no singular interest; for *any* block $\mathbf{a} = (a_0, a_1, \dots, a_{N-1})$ evidently $0 \times \mathbf{a} = \mathbf{0}$ and $1 \times \mathbf{a} = \mathbf{a}$, where $\mathbf{0} = (0, 0, \dots, 0)$ is a zero block formed of N zeros.

It is not difficult to verify that the operations of addition and multiplication by a number so defined satisfy all the basic rules of ordinary arithmetic operations. To express this fact in the language of modern algebra, we say that a collection of all possible N -tuple sequences $\mathbf{a} = (a_0, a_1, \dots, a_{N-1})$ of 0's and 1's forms a *vector space* (precise definition of a vector space is given in Appendix II).

On the other hand, the fact that the operation of adding two sequences has by itself (i.e., unrelated to multiplication by a number) the conventional properties of arithmetic operation of addition can be expressed by saying that a collection of sequences $\mathbf{a} = (a_0, a_1, \dots, a_{N-1})$ is a *group* under the addition operation introduced above (the definition of group is given in Appendix II). A code (i.e., a definite collection of code words, each of which is a 'block', i.e., an N -tuple sequence of 0's and 1's) is called *linear*, if its code words form a *linear subspace* of the common vector space of all such 'blocks', implying that the sum of any two code words of a linear code and also the product of a code word by a number z must be a code word.† A code is called a *group code* if its code words form a *subgroup* of a common group of sequences $(a_0, a_1, \dots, a_{N-1})$; in the binary case considered here, this again means just that the sum of any two code words and a 'zero block' $(0, 0, \dots, 0)$ must be a code word (see also Appendix II). Thus, it is seen that in the case of a binary channel (i.e., the case in which only two elementary signals are used), both the terms linear code and group code mean one and the same thing.††

Consider now an arbitrary (not necessarily systematic) parity-check code, whose code words coincide with a collection of sequences $\mathbf{a} = (a_0, a_1, \dots, a_{N-1})$ such that relations (1') are satisfied for them. In the first place, it is obvious that if $(a_0, a_1, \dots, a_{N-1})$ is a block of 0's alone, then relations (1') are necessarily satisfied; hence, the zero block $(0, 0, \dots, 0)$ is surely a code word of the considered code. Moreover, if the blocks

$$\mathbf{a} = (a_0, a_1, \dots, a_{N-1})$$

and

$$\mathbf{a}' = (a'_0, a'_1, \dots, a'_{N-1})$$

are both code words (i.e., all K relations (1') are satisfied for both of them), then adding to each other the first, second, \dots , last of these relations for \mathbf{a} and \mathbf{a}' it is verified that

$$\mathbf{a} + \mathbf{a}' = (a_0 + a'_0, a_1 + a'_1, \dots, a_{N-1} + a'_{N-1})$$

†It is clear that for the case considered in which only two elementary signals are admissible the condition related to multiplication by a number is rather trivial; it means only that a sequence $(0, 0, \dots, 0)$ of N 0's must be a code word. However, in the case of more than two elementary signals, the indicated condition turns out to be sufficiently important.

††In the more general case of a communication channel with m elementary signals, these two notions are equivalent to each other if $m = p$ is a prime number, but the notion of linear code is only a particular case of the notion of group code if $m = p^k$, where p is a prime and $k > 1$ (see footnote†† on p. 317). Finally, if m is not an integral power of some prime number, then neither of these notions can in general be defined,

also satisfies all relations (1'), i.e., is also a code word. This implies that every parity-check code is simultaneously also a linear (group) code. On the other hand, in algebra it is shown that any linear subspace of a vector space of sequences $\mathbf{a} = (a_0, a_1, \dots, a_{N-1})$ can be defined by a certain collection of relations of the form (1') (see Appendix II). Consequently, the class of linear (or group) codes for a binary channel coincides precisely with the class of parity-check codes; this fact provides the justification for also calling parity-check codes, the linear codes or group codes.

Let us revert to the consideration of general parity-check codes. It has already been remarked above that every such code can be represented in the form of a systematic code (satisfying relations of the form (1)); therefore, we shall deal here mainly with codes of the latter type. Such a code is defined by matrix (2), which is usually called the *check matrix* of a code†; for convenience let us denote it by a single letter B . If $\mathbf{a} = (a_0, a_1, \dots, a_{N-1})$ is one of the code words of our code, the validity of relations (1) for it is conveniently represented by the single vector equation

$$B\mathbf{a} = \mathbf{0}. \quad (3)$$

The left-hand side of (3) serves as the symbolic notation of $N - M$ entries of the left-hand sides of equations of the form (1') obtained from (1) by transferring all left-hand sides to the right-hand sides; here $B\mathbf{a}$ is the product of the matrix B and the vector \mathbf{a} , understood in the sense of matrix calculus (which is further dealt with in Appendix II). Suppose that the code word $\mathbf{a} = (a_0, a_1, \dots, a_{N-1})$ is transmitted through a communication channel; as a result of distortion in the transmission process, a sequence $\mathbf{a}' = (a'_0, a'_1, \dots, a'_{N-1})$ other than the one transmitted is in general obtained at the output. Substitute the sequence \mathbf{a}' in the left-hand sides of equations (1') (understood, as usual, in the sense of 2-arithmetic) and denote by the symbol $B\mathbf{a}'$ the resultant $K = N - M$ numbers 0 and 1 (representing a K -term sequence $(s_M, s_{M+1}, \dots, s_{N-1})$). Since \mathbf{a}' is in general not a code word, $B\mathbf{a}' = \mathbf{s} = (s_M, s_{M+1}, \dots, s_{N-1})$ is not a zero sequence (i.e., it also contains 1's at certain places). The presence of these 1's obviously shows the distortion to have occurred during transmission; in the language used previously, each 1 implies the corresponding 'parity check' to have led to a negative result. Let

$$\mathbf{e} = (e_1, e_2, \dots, e_N) = (a'_1 - a_1, a'_2 - a_2, \dots, a'_{N-1} - a_{N-1})$$

be an N -term 'error block', containing 1's at places corresponding to the signals

†In the case of general (not necessarily systematic) parity check codes, a check matrix is obviously an arbitrary $(K \times N)$ -matrix of 0's and 1's (some examples of such general check matrices will be given later).

a , distorted during transmission and 0's at all the remaining places, so that

$$e = a' - a = a' + a$$

(recall that in 2-arithmetic $a - b = a + b$). It is clear that by (3), we have

$$Be = B(a' - a) = Ba'.$$

Consequently,

$$Be = s. \quad (4)$$

Unfortunately, in general, there exist *many* sequences $e = (e_0, e_1, \dots, e_{N-1})$ that satisfy $N - M$ relations (4). Therefore, starting from (4) it is still not possible to reconstruct the 'error block' e (and hence also the transmitted sequence $a = a' - e = a' + e$). When decoding a parity-check code, it is usually assumed that the probability of distortion in transmission of each signal is smaller than the probability of correct transmission. In agreement with this the following *decoding rule* is set up: *as an error block e is taken that sequence satisfying equation (4) which contains the least number of 1's, i.e., corresponds to the least possible number of distortions in transmission (if among the sequences satisfying (4) there are several sequences containing one and the same least number of 1's, then e is chosen randomly from them)*. This rule allows us to decode all N -term sequences of elementary signals received at the channel output, i.e., associates with all of them a definite code word $a = a' + e$ (condition (3) is obviously satisfied by a , i.e., a is in fact a code word). This code word a is then considered to have been transmitted over the communication channel.

The described method of decoding a parity-check code is appreciably simpler than the general method set forth on p. 293 (and based on the consideration of the groups \mathcal{B} corresponding to distinct code words). Nonetheless even this is not practically suitable: for large values of $K = N - M$, finding a sequence satisfying (4) that contains the least number of 1's turns out to be so tedious that even modern computers are unable to accomplish it within a tolerable time. Hence, the problem of developing sufficiently simple (i.e., attainable in practice) methods for finding the required block e is quite important; for the present it can be regarded to have been solved only for some particular cases of codes with highly special check matrix B structures.† However, even without this the existence of the theoretically sufficiently straightforward general decoding rule indicated above can be put to use for studying the general pro-

†One such particular case was studied by Gallager [199]. It relates to the matrix B with large values of N and $K = N - M$ which consists, roughly speaking, almost only of 0's (i.e., contains just a few 1's). Some other particular algebraic cases from algebra are described below.

perties of parity-check codes. Such a study was inaugurated by Slepian [214]. Later Elias [182] showed that in the case of a binary symmetric channel (and also in the case of a binary erasure channel corresponding to the scheme of Fig. 21, where $p = 0$), parity-check codes are in no way inferior to the best of all existing codes in the sense that *by means of parity-check codes information transmission can always be effected at a given rate $C_1 = Lc_1$ bits/unit time (less than the capacity $C = Lc$ of a channel) such that the probability of decoding error is smaller than any preassigned number $\epsilon > 0$* . Moreover, the magnitude of error probability attainable for a fixed transmission rate $C_1 = Lc_1$ bits/unit time, where $c_1 < c$, and for code words of fixed length N , does not exceed a_1^{-N} , where a_1 is a number depending on c_1 but always greater than unity. Thus, with increasing N the error probability decreases by the same rule as that applicable also in the case of a best arbitrary code. In addition, Elias has also demonstrated that if a parity-check code is chosen 'at random' (i.e., every element $b_{i,j}$ of the check matrix B is chosen by flipping a coin and assuming that $b_{i,j} = 0$ or $b_{i,j} = 1$ according as the coin comes up heads or tails), even then for a given channel the probability of a decoding error tends to zero as $N \rightarrow \infty$ (and $K = (1 - c_1)N$, so that $2^{N-K} = 2^{c_1 N}$). Moreover, the error probability here tends to zero not more slowly than the N th power of some number smaller than unity).†

The fact that for many communication channels encountered in actual practice a randomly chosen parity-check-code for large N turns out to be sufficiently good ('almost with a certainty') provides a great attraction for the use of such 'random parity-check codes'. In order to define such a code, it is necessary to choose randomly (and memorize) $MK = N^2 c_1 (1 - c_1)$ elements $b_{i,j}$, (where $i = M, M + 1, \dots, N - 1$, and $j = 0, 1, \dots, M - 1$) of the corresponding check matrix B . Since the number $N^2 c_1 (1 - c_1)$ with increasing N does not increase very rapidly (much more slowly than, for instance, $2^{c_1 N}$), such problem can be fully tackled by modern computers even for an N of the order of several hundreds. However, the procedure of decoding (i.e., determining with respect to the received sequence a' the corresponding 'error block' e), as noted previously, is extremely difficult in the case of an arbitrarily chosen parity-check code and this substantially hinders the use of 'random codes'. Nevertheless there exist definite promising approaches to the construction of practically 'good' coding and decoding methods, with inbuilt provision for the 'random' choice of some variables defining the code under consideration (by way of example, we may mention the so-called 'sequential decoding', that has been described, for example,

†Later Dobrushin [194] (while investigating arbitrary group codes) and Drygas [196] (while considering more particular linear codes) extended the results of Elias related to a binary symmetric channel to more general channels with $m = p^k$ elementary signals and $r = m$ (i.e., where the same signals are transmitted and received, provided that the corresponding probabilities $p_{A_j}(A_i)$ satisfy some specific symmetry conditions. However, for arbitrary channels all these results remain untrue (see [189], [200]).

in [11] and the review paper [195]). Since all these approaches are quite complicated, we shall not dwell upon them here and immediately pass on to the application of 'nonrandom' parity-check codes for detecting and correcting the transmission errors.

Denote the individual columns of check matrix B (the 'blocks' of $K = N - M$ digits 0 and 1) by $b_0, b_1, \dots, b_{M-1}, b_M, \dots, b_{N-1}$ (in the case of a systematic code the last K columns b_M, \dots, b_{N-1} obviously all contain a single 1 and $N - M - 1$ zeros each). Then, the matrix B can be set up in the form of a single row

$$B = (b_0, b_1, \dots, b_{M-1}, b_M, \dots, b_{N-1}).$$

As above, denote by $e = (e_0, e_1, \dots, e_{N-1})$ the 'error block' containing 1's at the places of those elementary signals of a transmitted code word that are distorted during transmission. The basic equation (4) can then be written in the form

$$e_0 b_0 + e_1 b_1 + \dots + e_{M-1} b_{M-1} + e_M b_M + \dots + e_{N-1} b_{N-1} = s, \quad (5)$$

where the addition is understood to be termwise addition (in the sense of 2-arithmetic) of the corresponding 'blocks' of length K . Thus, the 'block' s , which is obtained by replacing in the left-hand sides of equations (1') the transmitted signals a_0, a_1, \dots, a_{N-1} by the received signals $a'_0, a'_1, \dots, a'_{N-1}$ (and is used then for determining the existing errors), is equal to the sum of the columns of B corresponding to the signals distorted during transmission (i.e., corresponding to the value $e_i = 1$, the remaining signals correspond to the value $e_i = 0$, and hence the corresponding summands $e_i b_i$ reduce to 0). This implies, in particular, that a *single* error (i.e., a block e containing a single 1 and $N - 1$ zeros) corresponds to the block s coinciding with the particular column b_i of B ; the occurrence of no error, however, corresponds to a block $s = 0$ of $N - M$ zeros. Hence, in order that a parity-check code allow us to distinguish between the cases of no error and all those of a single error, it is necessary that all columns of the corresponding check matrix B be distinct and that none of them be equal to 0.

The total number of possible distinct K -term blocks $b = (b_M, b_{M+1}, \dots, b_{N-1})$ (i.e., distinct K -term sequences of 0's and 1's) is equal to the number of integers written in the binary number system by means of not more than K digits, i.e., to 2^K (similar to this, the number of not more than K -digit distinct numbers in the decimal number system is 10^K). Since a zero block $(0, 0, \dots, 0)$ is excluded here from the number of possible columns of matrix B , the number of possible columns turns out to be $2^K - 1$. Thus, we again arrive at the conclusion that *a parity-check code, correcting all single errors and containing K 'check signals', must be formed of code words whose length does not exceed $2^K - 1$* . For defining such a code it is required only to indicate the corresponding check matrix B , all of whose columns must be nonzero and distinct,

The codes obtained here naturally coincide with the Hamming codes referred to on p. 314. In the case in which $N = 2^K - 1$ it is appropriate to set up the corresponding check matrix B by choosing as its columns the binary notations (i.e., notations in binary number system) for all integers from 1 to $2^K - 1$, counted in ascending order. It is apparent that the code obtained here is indeed systematic (since it contains all possible columns of $K - 1$ zeros and a single 1), except that the 'check signals' here are not the last K signals but some other K signals. Thus, for instance, in the case in which $K = 4$, $N = 2^4 - 1 = 15$, the corresponding (4×15) -matrix B is set up appropriately in the form

$$B = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix}.$$

(Note that if we desire here to write all code words in a way similar to that on p. 311 for application to the case in which $K = 3$, $N = 7$, then we shall have to write $2^{11} = 2048$ fifteen-digit numbers!) For such a matrix B the role of 'check signals' is assumed by the first, second, fourth and eighth signals (since the columns containing three 0's and a single 1 correspond specifically to them); the other 11 signals are information signals. The block s is here zero when there is no transmission error, and in the case of a single error is equal to the corresponding column of B , i.e., it determines directly a binary number notation of the elementary signal that is distorted during transmission. Hence, it is seen that in this case it is extremely straightforward to accomplish the decoding (i.e., deciphering the received signals and correcting the errors in them).

A single-error-correcting code related to blocks of $N < 2^K - 1$ signals is easy to obtain by deleting in the corresponding check matrix B a certain number of 'superfluous' columns (which can be chosen arbitrarily out of those containing not less than two 1's). It may also be noted that the properties of the Hamming code can be sharpened further by adding to each code word an auxiliary $(K + 1)$ th 'check signal' a_N , which even allows us to detect (but not correct) all double errors. To do this, the only requirement is to choose the binary signal a_N such that it yields an even number when added to all the rest of signals, i.e., it satisfies the relation

$$a_0 + a_1 + \dots + a_{N-1} + a_N = 0.$$

(It is easy to comprehend that this corresponds to adding to the check matrix B first an additional last column of only 0's and then an additional last row of $N + 1$ 1's; as a result, the number of both columns and rows of B increases by one.) In such a case, the absence of any error again corresponds to a block s

of only 0's; in the case of a single error the first K digits of s represent the binary notation of some integer in the range from 0 to $2^K - 1$, and the last digit s_{K+1} is unity (since the sum of all received signals is necessarily odd here); finally, the presence of even a single 1 among the first K elements of s and its last element reducing to 0 indicate the presence of a double error. The Hamming code thus refined is proposed also in [203]; it is sometimes called an *extended Hamming code*.

We now pass on to codes correcting not only all *single* but also all *double* errors in a block of N signals. It is clear that when there is *no* transmission errors a block $s = Ba'$ of K elements is composed of only 0's: in the case of *one* error it is equal to the corresponding column of the check matrix B and in the case of *two* errors, to the sum of two corresponding columns of B (cf. relation (5) on p. 324). In order that all of these cases be distinguished at the channel output, all columns of B must be nonzero, different from each other, and such that the sum of any two of them differs from all other columns and from the rest pairwise sums of the columns. Following Sacks [213] we can undertake to construct a matrix satisfying all these conditions by means of a simple sorting out. With this object, we can choose the first column b_0 of B in an arbitrary manner (but such that it does not consist of only 0's). Then we take as b_1 an arbitrary nonzero block of K digits 0 and 1 distinct from b_0 , as b_2 a nonzero block distinct from b_0 , b_1 and $b_0 + b_1$, and as b_3 a nonzero block distinct from b_0 , b_1 and b_2 , as well as from the pair sums $b_0 + b_1$, $b_0 + b_2$, $b_1 + b_2$ and the triple sum $b_0 + b_1 + b_2$ (because in 2-arithmetic if $b_0 + b_1 + b_2 = b_3$, then

$$b_0 + b_1 = b_2 + b_3,$$

i.e., the errors in the first two code word signals cannot be distinguished here from the errors in the third and fourth signals), and so on. After the first i columns b_0, b_1, \dots, b_{i-1} are so chosen, the prescription for the choice of the $(i + 1)$ th column b_i is that this column

(a) not be a zero column;

(b) not be equal to any of the $i = \binom{i}{1}$ columns b_0, b_1, \dots, b_{i-1} already chosen;

(c) be equal to none of the $\binom{i}{2}$ pairwise sum of columns already chosen;

(d) be distinct from all $\binom{i}{3}$ sums of the three already chosen columns.

Obviously, the enumerated $1 + \binom{i}{1} + \binom{i}{2} + \binom{i}{3}$ conditions (a) — (d),

restricting the choice of column b_i , are not necessarily all distinct among themselves. (Thus, for instance, for $i > 5$ it is possible that

$$b_0 + b_1 + b_2 = b_3 + b_4 + b_5,$$

or that

$$b_1 + b_2 + b_3 = b_4 + b_5.)$$

However, since the number of all distinct columns (i.e., blocks of K digits 0 and 1) is equal to 2^K , hence if only

$$1 + \binom{i}{1} + \binom{i}{2} + \binom{i}{3} < 2^K,$$

then conditions (a) — (d) can surely be satisfied even in the least favourable case in which all columns and their combinations figuring in these conditions are distinct. Of the relations obtained here, the most restrictive is one applied to the last column b_{N-1} (since with increasing i the number of excluded combinations, with which a new column must not coincide, also increases). Hence, if only

$$2^K > 1 + \binom{N-1}{1} + \binom{N-1}{2} + \binom{N-1}{3},$$

i.e.,

$$K > \log \left[1 + \binom{N-1}{1} + \binom{N-1}{2} + \binom{N-1}{3} \right], \quad (****)$$

then a $(K \times N)$ check matrix B can certainly be chosen that yields a parity-check code correcting all single errors and all double errors in a block of N elementary signals.

The inequality obtained here is the *Varshamov-Gilbert inequality*, which was given on p. 316 without proof (for the case of an arbitrary number n of errors corrected by our code). It is clear that in the general case of an arbitrary n this inequality is proved exactly in the same way as in the case in which $n = 2$. The only requirement now is that each time the new column b_i must not be a null column, or equal to any of the previous columns, or equal to any of the sums of the two, three, . . . , $2n - 1$ preceding columns. This implies the general Varshamov-Gilbert inequality

$$K > \log \left[1 + \binom{N-1}{1} + \binom{N-1}{2} + \dots + \binom{n-1}{2n-1} \right]. \quad (***)$$

Let us again consider that $n = 2$. It is obvious that for small values of K and N it is possible to hope that all conditions imposed on the columns of matrix B can be directly verified, thus giving a construction of single-error-correcting and double-error-correcting code. This is, in fact, the method used on p. 314, where for the case in which $K = 4$ and $N = 5$ the selection procedure

was used to construct a parity-check code which permits us to correct all single and all double errors. The check matrix corresponding to this code obviously has the form

$$\begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

(Note that for $N = 5$ and $n = 2$ the Hamming inequality indicates that $K \geq 4$ necessarily; from the Varshamov-Gilbert inequality, however, it follows here that for $K \geq 4$ we can in fact construct a code correcting all single and double errors.) Although slightly more intricate, it is completely possible to verify the fact that for $K = 7$ and $N = 10$ all columns and pairwise sums of columns of the (7×10) -matrix

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{bmatrix}$$

are distinct from each other. Hence, the corresponding code (all of whose code words contain 3 information signals and 7 check signals) allows us to correct all single and double errors in blocks of 10 signals. (For $N = 10$, from the Hamming inequality it follows that $K \geq 6$ necessarily, and from the Varshamov-Gilbert inequality it follows that for $K \geq 8$ the code we are interested in can certainly be constructed.)

However, further increase in the values of K and N rapidly increases the unwieldiness of the described procedure for choosing matrix B and verifying the validity of the requisite conditions for the columns of this matrix. For example, in the case of an (8×15) -matrix B given later on p. 335, the problem of carrying out all necessary checks is hardly any different.

We shall now briefly sketch some fundamental principles of *algebraic coding theory*. This theory has played a central role in the development of general methods for constructing practical usable codes, which allow the detection and correction in a block of N signals of any number of errors not exceeding a given number n . So far we have considered a code as a collection of some code words, i.e., blocks $a = (a_0, a_1, \dots, a_{N-1})$ of N digits 0 and 1 (i.e., of N elements of the simplest binary algebraic field). It is clear that we can also associate with

every code word a *code polynomial* of power not higher than $N - 1$:

$$a(x) = a_0 + a_1x + a_2x^2 + \dots + a_{N-1}x^{N-1}$$

with the coefficients of our field. We may then consider a code as a collection of 'code polynomials' $a(x)$. All possible parity-check codes (i.e., all group codes) in such a case correspond to all possible collections of polynomials $a(x)$ such that the sum of any two polynomials belonging to our collection, and also the 'null polynomial' $0 = 0 + 0 \times x + \dots + 0 \times x^{N-1}$ necessarily belong to the same collection. There is an extensive class of quite simple collections of polynomials which obviously satisfy the two indicated conditions. This class consists of collections of all polynomials $a(x)$ of degree not greater than $N - 1$, which are divisible by a fixed polynomial $g(x) = g_0 + g_1x + \dots + g_Kx^K$ of degree $K < N - 1$, i.e., can be represented in the form

$$a(x) = c(x)g(x), \quad (6)$$

where $g(x)$ is a fixed polynomial and $c(x)$ is an arbitrary polynomial of degree not exceeding $N - K - 1$. Each such collection determines a definite parity-check code, which we call a *code generated by the polynomial* $g(x)$; $g(x)$ itself in this case is called the *generator polynomial*, or simply *generator* of our code. It is clear that coefficients g_0 and g_K of the generator polynomial must be different from 0 for all polynomial generated codes. In fact, if $g_0 = 0$, then the first coefficient a_0 of all the code words (6) is also equal to 0, i.e., the first signal of the block $\mathbf{a} = (a_0, a_1, \dots, a_{N-1})$ contains no information. If $g_K = 0$, then we must just consider the generator polynomial $g(x)$ as the polynomial of degree $K - 1$.

In the case of polynomial generated codes, the bonus from the generator polynomials is a highly compact method of determining the corresponding code, which uniquely defines all of its characteristics (in particular, the collection of all code words \mathbf{a} and the corresponding check matrix \mathbf{B}). If an arbitrary code polynomial $a(x)$ of such a code is expressed in the form

$$a(x) = a_0 + a_1x + \dots + a_{K-1}x^{K-1} + a_Kx^K + a_{K+1}x^{K+1} + \dots + a_{N-1}x^{N-1},$$

then it is apparent here that the last $M = N - K$ coefficients $a_K, a_{K+1}, \dots, a_{N-1}$ can be chosen arbitrarily, and the first K coefficients a_0, a_1, \dots, a_{K-1} are then uniquely determined by the condition of the divisibility of $a(x)$ by $g(x)$. (Specifically, since in 2-arithmetic $r(x) = -r(x)$, the polynomial $a_0 + a_1x + \dots + a_{K-1}x^{K-1}$ must coincide with the remainder after dividing $a_Kx^K + a_{K+1}x^{K+1} + \dots + a_{N-1}x^{N-1}$ by $g(x)$.) From this it is seen that the last $N - K$ signals $a_K, a_{K+1}, \dots, a_{N-1}$ in the given case correspond to information signals and the first K signals a_0, a_1, \dots, a_{K-1} to check signals. Hence the considered code is

a block $(N, N - K)$ -code and the total number of code words here is 2^{N-K} . A block $\mathbf{a}' = (a'_0, a'_1, \dots, a'_{N-1})$ received at the channel output corresponds to the polynomial

$$a'(x) = a'_0 + a'_1x + \dots + a'_{N-1}x^{N-1},$$

which differs from the 'transmitted polynomial' $a(x)$ by the 'error polynomial'

$$e(x) = e_0 + e_1x + \dots + e_{N-1}x^{N-1},$$

where, as previously, $e_i = a'_i - a_i$ (i.e., $e_i = 1$ if the i th signal is distorted during transmission, and $e_i = 0$ if it is received correctly). Due to the presence of an additional 'error polynomial' $e(x)$, the polynomial $a'(x)$ is in general not evenly divisible by $g(x)$. The nonzero remainder $r(x)$ resulting from the division of $a'(x)$ by $g(x)$ (obviously equal to the remainder after dividing $e(x)$ by $g(x)$) is also an indicator of the occurrence of distortions during transmission; this remainder contains all information about errors transmitted to the receiving end. (The remainder $r(x)$ is in this respect completely analogous to the block $\mathbf{s} = \mathbf{B}\mathbf{a}'$ that we dealt with in the matrix description of an arbitrary parity-check code.)

The foregoing discussion shows that the collection of all detectable error blocks $\mathbf{e} = (e_0, e_1, \dots, e_{N-1})$ can be described very easily in the case of a polynomial generated code. In fact, it follows from above-stated results that block \mathbf{e} is detectable if and only if the corresponding error polynomial $e(x)$ yields nonzero remainder when divided by $g(x)$. The correction of error is also often possible when a polynomial generated code is used. To explain this, let us first of all remark that two code polynomials $a(x)$ and $a_1(x)$ of a polynomial generated code cannot differ in less than two coefficients. This is clear since the difference of two code words must be divisible by $g(x)$ without remainder and if $a(x)$ and $a_1(x)$ differ in only one coefficient, then their difference is proportional to x^i and, therefore, it cannot be divisible by any polynomial $g(x) \neq 1$ with $g_0 \neq 0$. Moreover, if $g(x)$ does not coincide with a divisor of the polynomial of the form $x^L - 1$, where $L < N$, then two code polynomials $a(x)$ and $a_1(x)$ cannot differ in less than three coefficients. In fact, if $a(x)$ and $a_1(x)$ differ in only two coefficients, then

$$a(x) - a_1(x) = x^i - x^j = x^j(x^{i-j} - 1), \quad \text{where } 0 \leq j < i \leq N - 1$$

(let us remind that $x^i + x^j = x^i - x^j$ in 2-arithmetic). The last relation shows that $g(x)$ must be a divisor of $x^{i-j} - 1$. Therefore, if $g(x)$ is not a divisor of any polynomial $x^L - 1$, $L = 1, 2, \dots, N - 1$, then the smallest number d of distinct elementary signals is not less than 3 for any pair of code words, i.e., the code permits to correct any single error (see p. 309). Similarly, any two code polynomials $a(x)$ and $a_1(x)$ will differ in more than three coefficients, if and only

if generator polynomial $g(x)$ is not a divisor of any polynomial of the form

$$\text{either } x^i + x^j = x^i - x^j = x^j(x^{i-j} - 1), \text{ or } x^i + x^j + x^k,$$

and so on.

In algebraic coding theory, attention is mainly focused not on general parity-check codes, nor even on more special arbitrary polynomial generated codes, but on some particular classes of such polynomial generated codes, having a singularly simple algebraic structure that appreciably facilitates obtaining a practically convenient coding and decoding procedure. Of these particular classes, the most important one is that of cyclic codes. A parity-check code is called *cyclic*, if for each of its code words $\mathbf{a} = (a_0, a_1, a_2, \dots, a_{N-1})$, a block $(a_{N-1}, a_0, a_1, \dots, a_{N-2})$, which is obtained by *shifting a cyclically*, is also a code word. It is clear that in such a case a block $(a_{N-i}, a_{N-i+1}, \dots, a_{N-1}, a_0, \dots, a_{i-1})$ obtained by performing an i -multiple 'cyclic shifting' of \mathbf{a} is also a code word for every $i = 1, 2, \dots, N - 1$.

An important property of cyclic codes is that they are all generated by polynomials and it is quite simple to characterize the class of generating polynomials $g(x)$ corresponding to them. In fact, let us first assume that we are concerned with the code generated by the polynomial $g(x)$ (i.e., with a collection of code polynomials $a(x)$ of the form (6)). Suppose that

$$a_1(x) = a_{N-1} + a_0x + a_1x^2 + \dots + a_{N-2}x^{N-1}$$

is a polynomial corresponding to a shifted block $(a_{N-1}, a_0, a_1, \dots, a_{N-2})$. Since

$$\begin{aligned} a_1(x) &= x(a_0 + a_1x + \dots + a_{N-1}x^{N-1}) - a_{N-1}(x^N - 1) \\ &= xa(x) - a_{N-1}(x^N - 1), \end{aligned} \quad (7)$$

where, as usual, $a(x) = a_0 + a_1x + \dots + a_{N-1}x^{N-1}$, it is clear that for the general case in which $a_{N-1} \neq 0$, $a_1(x)$ is a code polynomial simultaneously with $a(x)$ (i.e., is evenly divisible by $g(x)$) if and only if $g(x)$ is a factor of $x^N - 1$.† Thus, a code generated by a polynomial $g(x)$ is cyclic in that (and only that) case in which $g(x)$ is a factor of the polynomial $x^N - 1$.

Consider now an absolutely arbitrary cyclic code, and suppose that $a(x)$ is a code polynomial corresponding to it. Then, from equation (7) it follows directly that, together with $a(x)$, in the collection of code polynomials of our code there necessarily occurs also a remainder after dividing the polynomial $xa(x)$ by $x^N - 1$. But then it is clear that in the collection of code polynomials there are also remainders after dividing the polynomials $x \times xa(x) = x^2a(x)$, $x \times x^2a(x) = x^3a(x)$, \dots by $x^N - 1$, i.e., remainders after dividing all possible products

†Such polynomials $g(x)$ are called *cyclotomic* in algebra; the case for which the coefficients of $g(x)$ are ordinary real numbers was studied extensively by the great German mathematician Carl Friedrich Gauss at the turn of the nineteenth century.

$x^n a(x)$ by $x^N - 1$, where n is any nonnegative integer. Since, furthermore, the sum of any code polynomials is always a code polynomial also, it follows from our assertions that, *together with $a(x)$, all remainders after dividing polynomials of the form $b(x)a(x)$ by $x^N - 1$ are also code polynomials, where $b(x) = b_0 + b_1x + \dots + b_nx^n$ is an arbitrary polynomial with coefficients from our two-element field (i.e., either 0 or 1).*

A collection of all possible polynomials of degree not greater than $N - 1$ can be considered as a collection of all possible remainders resulting from the division of polynomials of any degree by $x^N - 1$. Then, the property enunciated above of a collection of code polynomials $a(x)$ of an arbitrary cyclic code can be stated as follows in the language of general algebra: such a collection of code polynomials is an *ideal* in a set of all remainders after dividing the polynomials by $x^N - 1$ (see Appendix II, where a general definition of an ideal is given, and a particular case of this notion required by us is also considered). In the following, we shall not use the general definition of an ideal; we need only the following straightforward algebraic theorem (which the reader, if desired, may accept on faith, but may also acquaint himself with its proof from Appendix II): *any ideal in a set of remainders from the division of arbitrary polynomials by some fixed polynomial $f(x)$ of degree N coincides with a collection of polynomials of the form $c(x)g(x)$, where $g(x)$ is some fixed factor of the polynomial $f(x)$ and the degree of $c(x)g(x)$ is not greater than $N - 1$.* This algebraic theorem evidently implies that *every cyclic code is generated by some factor $g(x)$ of the polynomial $x^N - 1$.*

Suppose that $g(x)$ is a factor of $x^N - 1$ and hence

$$x^N - 1 = g(x)h(x).$$

In such a case it is easy to show that the code polynomials of a cyclic code with generator polynomial $g(x)$ are such polynomials $a(x)$ of degree not exceeding $N - 1$, for which $a(x)h(x)$ is evenly divisible by $x^N - 1$. In fact, if $a(x) = c(x)g(x)$, then it is obvious that

$$a(x)h(x) = c(x)g(x)h(x) = c(x)(x^N - 1)$$

is evenly divisible by $x^N - 1$; conversely, if $a(x)h(x) = b(x)(x^N - 1)$ is evenly divisible by $x^N - 1$, then it is clear that $a(x) = b(x)g(x)$. The indicated characteristic of $a(x)$ greatly facilitates to check the occurrence of transmission errors: if

$$a'(x) = a(x) + e(x),$$

where $e(x) \neq 0$, then $a'(x)h(x)$ is in general not divisible by $x^N - 1$. It is also easy to see that all information concerning the occurrence of errors (i.e., about the polynomial $e(x)$), available at the channel output is contained in the remain-

der from the division of $a'(x)h(x)$ by $x^N - 1$. (Note that the division of an arbitrary polynomial $d(x)$ by $x^N - 1$ is most easy to perform: for this the only requirement is to replace in $d(x)$ all powers x^M , where $M \geq N$, by the power x^m , where m is the remainder after dividing M by N .) Hence, in decoding a cyclic code, a vital role is played by the polynomial $h(x)$, which we agree to call the *check polynomial of a cyclic code*. In fact, the polynomial $a'(x)$ received at the channel output should first be multiplied by the check polynomial $h(x)$, and then the remainder resulting from the division of this product by $x^N - 1$ uniquely determines the deciphering of the received message (i.e., the choice of the 'most probable error polynomial' $e(x)$).

Cyclic codes form a special sub-class of parity-check codes, whose general characteristics have so far not been studied to a great extent. Thus, for instance, if we confine ourselves to the use of cyclic codes alone, then it is not known that we may or may not attain information transmission over the simplest binary symmetric channel at a given rate less than $C = Lc$ bits/unit time and with error probability as small as desired. Moreover, even it is not known whether or not the transmission can be effected at least at a rate different from zero and with error probability as small as desired.[†] However, the great advantage of cyclic codes lies in the fact that here we may develop some relatively uncomplicated algebraic decoding methods, which allow us in many cases to accomplish this decoding in relatively short time (see, for example, references cited on pp. 305-306 and also rather the advanced book [207], especially devoted to this problem).

The application of cyclic codes is exceptionally fruitful for correcting all errors whose number does not exceed a given n in an N -term block. According to the foregoing discussion related to arbitrary polynomial generated codes, in order that it be possible to correct all *single* errors by using a cyclic code generated by the polynomial $g(x)$, the only requirement is that *none* of the binomials

$$x^j - x^i = x^i(x^{j-i} - 1), \text{ where } i < N, j < N \text{ and } j > i,$$

be divisible by $g(x)$. The polynomials $g(x)$ with the prescribed properties, which are factors of $x^N - 1$ (i.e., correspond to cyclic codes), always exist and have been well studied for all $N = 2^K - 1$. Hence all the Hamming codes with $N = 2^K - 1$ can easily be put into the form of cyclic codes. In the particular, it is easy to verify that for the case in which $K = 3$, $N = 7$ (considered on pp. 310-311) the generator polynomial $g(x)$ and the check polynomial $h(x)$ can be

[†]Recall, as remarked on pp. 273-274, that until the appearance of Shannon's work [21] the impossibility of such transmission looked probable even for the case in which quite arbitrary codes are used. It is now known that for arbitrary codes the situation is entirely different, and the same is true for general parity-check codes. However, in relation to more special cyclic codes alone the impossibility indicated has not yet been ruled out.

chosen in the form

$$g(x) = x^3 + x + 1, \quad h(x) = x^4 + x^2 + x + 1$$

(direct multiplication shows that $g(x)h(x) = x^7 - 1$, as it should). Moreover, for the case in which $K = 4$, $N = 15$ (considered on p. 325), it is possible to set

$$g(x) = x^4 + x + 1, \quad h(x) = x^{11} + x^8 + x^7 + x^5 + x^3 + x^2 + x + 1$$

(here $g(x)h(x) = x^{15} - 1$).

Analogously, for *double-error-correcting* codes that permit to correct all single and double errors, all monomials x^i , binomials $x^i + x^j$, trinomials $x^i + x^j + x^k$ and quadrinomials $x^i + x^j + x^k + x^l$, where $i, j, k, l < N$, when divided by $g(x)$ must yield distinct remainders, and so on. It is clear that the problems arising here are specifically algebraic in their character; their solution turns out to be sufficiently involved, however.

A general method for the construction of cyclic codes, capable of correcting any number of errors less than n in a block of length $N = 2^K - 1$ and having a check matrix with nK rows and N columns (i.e., containing not more than nK check signals in a block of $N = 2^K - 1$ signals[†]), was indicated only in 1959 by Hocquenghem [204] and independently by Bose and Chaudhuri [192] in 1960.^{††} The Bose-Chaudhuri-Hocquenghem construction is not very complicated, but it is based on some relatively advanced algebraic concepts and results. These concepts and results can be found in Appendix II at the end of the book and the construction indicated will be described at the end of this section. The reader may well skip over this matter in case he is not interested in this algebraic construction.) Here we restrict ourselves to two examples of the error-correcting Bose-Chaudhuri-Hocquenghem codes. Both these examples relate to the case, in which $K = 4$, $N = 2^4 - 1 = 15$. The Hamming code correcting all *single* errors, which corresponds to these values of K and N , is defined by the check matrix written on p. 325. In the case of a double-error-correcting code, the check matrix can be represented as an (8×15) -matrix of the form

[†]Since the corresponding code is not systematic, hence from the fact that the check matrix contains nK rows it can only be inferred that the actual number of check signals here is not greater than nK (see p. 319).

^{††}Generally speaking, besides the simplest (the so called primitive) Bose-Chaudhuri-Hocquenghem codes correcting a given number of errors in a block of $N = 2^K - 1$ signals, there exist also non-primitive codes of the same type, for which the block-length N is an odd number not representable in the form $2^K - 1$. We shall not consider these last codes in this book (see, however, footnote[†] on p. 341),

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

This matrix is quite cumbersome; hence it is much more convenient to set the corresponding code with the aid of its generator polynomial

$$\begin{aligned} g(x) &= (x^4 + x + 1)(x^4 + x^3 + x^2 + x + 1) \\ &= x^8 + x^7 + x^6 + x^4 + 1, \end{aligned}$$

or its check polynomial

$$\begin{aligned} h(x) &= (x + 1)(x^2 + x + 1)(x^4 + x^3 + 1) \\ &= x^7 + x^6 + x^4 + 1 \end{aligned}$$

(it is easy to verify that, indeed, $g(x)h(x) = x^{15} - 1$). Note that the code under consideration consists of code words of length 15, involving 7 information and 8 check signals. By virtue of the Hamming inequality (*) on p. 315, we can say that for $N = 15$ a code correcting all single errors and all double errors cannot contain less than 7 check signals; here the Varshamov-Gilbert inequality (***) on p. 327 shows that such a code can certainly be constructed if $K = 9$.

If we now wish to construct a code, correcting in a block of 15 signals all *single*, *double* and *triple* errors, then the check matrix of such a Bose-Chaudhuri-Hocquenghem code has $3K = 12$ rows (and, as previously, 15 columns). The generator polynomial of the code we are interested in assumes the relatively simple form

$$\begin{aligned} g(x) &= (x^2 + x + 1)(x^4 + x + 1)(x^4 + x^3 + x^2 + x + 1) \\ &= x^{10} + x^9 + x^5 + x^4 + x^2 + x + 1, \end{aligned}$$

and its check polynomial is given by

$$h(x) = (x + 1)(x^4 + x^3 + 1) = x^5 + x^3 + x + 1$$

(here again we have $g(x)h(x) = x^{15} - 1$). The (12×15) -check matrix of our code is

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 \end{bmatrix}.$$

Note that although this matrix has 12 rows, the number of 'check signals' associated with the corresponding code is 10. This is immediate from the fact that the generator polynomial $g(x)$ is in this case a polynomial to the power 10.[†] Thus, in this use of the code under consideration every set of five 'information signals' is supplemented by ten 'check signals'. It is only then that in a sequence of 15 signals received at the channel output it is possible to detect and correct without exception all single, double and triple errors. It is also routine to see that the correction of all such errors in a block of 15 signals can in no way be achieved using less than 10 'check signals'. The fact is immediate from the Hamming inequality (the Varshamov-Gilbert inequality here shows that the code we require can surely be constructed if 12 or more 'check signals' are employed).

The data on the number of 'information' and 'check' signals for many distinct Bose-Chaudhuri-Hocquenghem codes can be found in various books on coding and information theory (see, for instance [212, Chapter 9], [190, Chapters 7 and 12]). By the results deduced in [212] all codes of this type with $N \leq 15$ and even those with N arbitrary but $n = 2$ are *optimal* in the sense that there does not exist a code with the same length of N 'blocks' and the same total number of code words S (i.e., with the same information rate $v = (L/N) \log S$ bits/unit time), leading to the lower error probability when it is used for transmission over a binary symmetric channel (see p. 342). When $N = 1023$ ($= 2^{10} - 1$) the number of 'check signals' for different values of n turns out to be quite close to the corresponding Varshamov-Gilbert bound. But for still larger N this must approach more closely to the Hamming lower bound and not to the

[†]The same conclusion in the considered case can be drawn by starting from the very form of the check matrix. Since its third row from the bottom consists of 0's alone and the two rows following it are identical, it is clear that the code is not affected if of the last three rows we retain only one (the last or penultimate) row.

Varshamov-Gilbert upper bound. In fact, if we use the upper bound of the binomial coefficient $\binom{N}{n}$ given by the inequality (**) on p. 165 and a similar lower bound of these coefficients (or even simply substitute into the exact formula $\binom{N}{n} = N!/n!(N-n)!$ the approximate values of the factorials $N!$ and $(N-n)!$ for large N , available in many advanced mathematics texts), it is routine to show that for very large N the general Hamming inequality assumes the form

$$2^K \geq AN^n, \quad \text{i.e., } K \geq n \log N + A_1,$$

where K is the number of check signals, n the maximum number of errors to be corrected, and A and $A_1 = \log A$ are some numbers (A is positive, but A_1 may possibly be negative) depending on n but not on N . Similarly, the Varshamov-Gilbert inequality in the case of large N allows us to conclude that if

$$2^K > BN^{2n}, \quad \text{i.e., } K > 2n \log N + B_1,$$

where B and $B_1 = \log B$ are other numbers depending on n (but not on N), then there does exist a code that enables us to correct any number of transmission errors not exceeding n in a block of N signals. In the case of Bose-Chaudhuri-Hocquenghem codes with $N = 2^{K_1} - 1$ (so that $K_1 \approx \log N$) the number K of check signals, as indicated above, does not exceed $nK_1 \approx n \log N$; hence for large values of N the number of check signals in these codes is always close to the corresponding Hamming lower bound. In this sense, these codes are close to the best possible codes with regard to their capability to correct a given fixed number of errors in very lengthy blocks.

Obviously, the choice of quite lengthy code words (i.e., extremely large N) is not advantageous if the codes correct only a fixed number n of errors, since with increasing N the probability of the emergence of errors more than n in a block of length N sharply increases. Hence, when N increases it is natural for the value of n also to increase simultaneously. However, if n increases proportionally to N , then with the increase of N , as has been shown, the information transmission rate decreases at the same time (see, for example, [212], Chap. 9). The most important problem, however, is not that of the optimal choice of the values of N and n but that of the method of decoding the obtained codes when N is large; specifically, the difficulty in decoding is the foremost constraint that restricts the opportunities for the choice of code parameters that will ensure both a low probability of error and a high transmission rate. In relation to Bose-Chaudhuri-Hocquenghem codes a whole series of special decoding methods have been developed that allow one to accomplish it effectively up to a length N of code words of an order of many hundred or even a few thousands. It is,

however, not possible here to dwell upon these methods; in this connection, we can only refer the reader to other (sufficiently advanced) works on information theory and coding referred to on pp. 305-306 (see also [198] and [207]). Several other interesting and practically useful codes have also been described in these works, but these have not been considered in the present text.

As in the foregoing, we shall consider only the case of a binary communication channel (using two elementary signals), and a code shall be understood as some collection of code words, i.e., sequences $a = (a_0, a_1, \dots, a_{N-1})$ of N digits 0 and 1. In the study of the error-correcting codes, an important role is played by the *Hamming distance* $|b - a|_H$, between two sequences $b = (b_0, b_1, \dots, b_{N-1})$ and $a = (a_0, a_1, \dots, a_{N-1})$, which by definition is equal to the number of digits a_i such that $b_i \neq a_i$ (i.e., the number of 1's between the difference $b_i - a_i$, understood in the sense of 2-arithmetic). The Hamming distance shares many characteristics of the usual geometric distance (see, for example, Appendix II.) It coincides with the number of distortions of individual signals to be transmitted, leading to the result that the transmitted sequence a is received as the sequence b at the channel output. Clearly, the larger the Hamming distance between individual code words, the smaller is the probability of confusing them at the receiving end, i.e., other conditions remaining the same, the better is the code to be used. Hence an important characteristic of a code is the *code distance*

$$D = \min |a^{(i)} - a^{(j)}|_H$$

associated with it, the Hamming distance between the 'closest' distinct code words of a given code. It is apparent that in the case of a code that allows us to correct any number of errors not exceeding n , the Hamming distance between all pairs of code words $a^{(i)}$ and $a^{(j)}$ must be greater than $2n$ (see p. 309). This implies that $D \geq 2n + 1$ here, D being the code distance of our code. Conversely, if $D \geq 2n + 1$, then by agreeing to decode as code word $a^{(i)}$ any received sequence b , belonging to the *Hamming sphere* of radius n with centre $a^{(i)}$ (i.e., all b such that $|b - a^{(i)}|_H \leq n$), we are sure to correct any number of transmission errors not greater than n . Thus, *a code is capable of correcting any number of transmission errors not greater than n if and only if its code distance D is not less than $2n + 1$* . Similarly, it is easy to show that *if the code distance D is not less than $2n$, then the code allows us to correct any number of errors not exceeding $n - 1$ and, in addition, to detect the occurrence of n errors* (but in the latter case it may not also be possible to correct precisely these n errors).†

It is clear that the 'volume' V_n of the Hamming sphere of radius n , i.e., the number of 'points' $b = (b_0, b_1, \dots, b_{N-1})$ belonging to this sphere with centre at an arbitrary 'point' $a = (a_0, a_1, \dots, a_{N-1})$ is defined by the formula

$$V = 1 + \binom{N}{1} + \binom{N}{2} + \dots + \binom{N}{n}.$$

Since the total number of all N -term sequences is 2^N , it is immediate that the number S of distinct code words of length N appearing in a code that allows us to correct any number

†It ought to be borne in mind that the code distance D does not define the total capability of the code to correct transmission errors. Thus, say, if $D = 2n$, then frequently for many (but not for all) transmitted words $a^{(i)}$ the codes nonetheless permit us to correct transmission errors even when they considerably exceed n errors.

of errors not exceeding n must satisfy the condition

$$S \leq \frac{2^N}{1 + \binom{N}{1} + \dots + \binom{N}{n}}. \quad (8)$$

This simple condition, giving an upper bound on the possible number S of code words (and hence also the maximum possible information transmission rate $v = (L/N) \log S$ bits/unit time), is called the *Hamming upper bound on the number of code words*. In the particular case of parity-check codes (i.e., differently, linear or group codes), it coincides with the lower Hamming bound on the number of check signals considered on p. 315: in fact, for an (N, M) -parity-check code the number S of code words is given by

$$2^M = \frac{2^N}{2^K},$$

and hence condition (8) coincides here exactly with the Hamming inequality. Note, however, that condition (8), in contrast to the Hamming inequality for the number K , applies to *any* code and not to parity-check codes alone.

A code having the property that the left-hand and right-hand sides of (8) coincide with each other is called a *perfect code* (or, sometimes, a *densely packed code*). Perfect codes are remarkable because in practically all respects they are optimal (i.e., the best). It is seen, for example, that among codes of a given length N , correcting a given number n of errors, *the largest number S of code words* (i.e., *the largest information transmission rate*) corresponds to perfect codes. Moreover, in the case of perfect parity-check codes correcting a specified number of errors, *the number of check signals K is the least possible*. Now assume that our code is employed for the transmission of information over a binary symmetric channel; here an extremely important characteristic of the quality of transmission is the *mean probability of decoding error*

$$Q = \frac{Q_1 + Q_2 + \dots + Q_S}{S},$$

where S is the total number of code words of the code and Q_i is the probability that the transmitted i th code word $a^{(i)}$ will be decoded in error at the channel output. Now suppose that $m_k^{(i)}$ is the number of sequences b , which are at Hamming distance k apart from the i th code word $a^{(i)}$ and are to be decoded as $a^{(i)}$ at channel output. Since in the case of the transmission of a sequence $a^{(i)}$ through a binary symmetric channel the probability of receiving any such sequence b is $p^k(1-p)^{N-k}$, the probability of decoding accurately the transmitted sequence $a^{(i)}$ is the sum

$$m_0^{(i)}(1-p)^N + m_1^{(i)}p(1-p)^{N-1} + \dots + m_k^{(i)}p^k(1-p)^{N-k} + \dots$$

Hence, it is seen that the mean probability of decoding error is

$$Q = 1 - \frac{1}{S} [m_0(1-p)^N + m_1p(1-p)^{N-1} + \dots + m_kp^k(1-p)^{N-k} + \dots],$$

where $m_k = m_k^{(1)} + m_k^{(2)} + \dots + m_k^{(S)}$ is the total number of sequences b at the Hamming distance k away from some code word $a^{(i)}$ and to be decoded as this $a^{(i)}$ (so that

$$m_0 + m_1 + \dots + m_k + \dots = 2^N).$$

But the total number of sequences of length N at a given Hamming distance k from a fixed sequence $a^{(i)}$ is $\binom{N}{k}$. Hence, for a code consisting of S code words of length N , we have

$$m_0 \leq S, m_1 \leq S \binom{N}{1}, \dots, m_k \leq S \binom{N}{k}, \dots$$

Suppose now that n is the largest integer such that

$$S + S \binom{N}{1} + \dots + S \binom{N}{n} < 2^N,$$

but

$$S + S \binom{N}{1} + \dots + S \binom{N}{n} + S \binom{N}{n+1} > 2^N,$$

so that

$$2^N - \left(S + S \binom{N}{1} + \dots + S \binom{N}{n} \right) = T < S \binom{N}{n+1}.$$

Then, if $m_0 = S, m_1 = S \binom{N}{1}, \dots, m_n = S \binom{N}{n}$, we have $m_{n+1} \leq T$. As usual, it is assumed here that $p < \frac{1}{2}$; then, the probability $p^k(1-p)^{N-k}$ decreases with the increase of k and hence the case in which $m_0 = S, m_1 = S \binom{N}{1}, \dots, m_n = S \binom{N}{n}, m_{n+1} = T$, is the most favourable, i.e., gives the *least* mean error Q . Consequently,

$$\begin{aligned} Q \geq 1 - \left[(1-p)^N + \binom{N}{1} p(1-p)^{N-1} \right. \\ \left. + \dots + \binom{N}{n} p^n(1-p)^{N-n} + \frac{T}{S} p^{n+1}(1-p)^{N-n-1} \right]. \end{aligned} \quad (9)$$

The estimate (9) of the least possible mean probability of decoding error for a code with fixed values of N and S (used for transmission over a binary symmetric channel with a given value of the probability p of the distortion of signals) is called the *Hamming lower bound on the mean probability of error*. For the perfect codes with decoding rule stating that all received N -term sequences separated from a code word $a^{(i)}$ by a Hamming distance not exceeding n are decoded as $a^{(i)}$, inequality (9) obviously turns into an equality (T being equivalent to 0 here). Hence, it is seen that for such codes the *mean probability of error is smaller than for any other code* with the same values of N and S .

Perfect codes have an extremely simple geometric meaning (in geometry that uses the Hamming distance instead of the usual distance): they correspond to the cases in which a collection of all possible 'points' $b = (b_0, b_1, \dots, b_{N-1})$ can be partitioned into a finite number of 'Hamming spheres' of a certain radius n , mutually disjoint but filling in their totality the entire 'space' (consisting of 2^N points), and the centres of these 'spheres' are taken as code words (hence the name 'densely packed code'). Their main deficiency is that only a few such codes are available, existing only for certain exclusive values of N and S . The simplest perfect code is a trivial code consisting of two code words $(0, 0, \dots, 0)$ and $(1, 1, \dots, 1)$, each of

which is composed of an odd number $N = 2n + 1$ of the same digits. For such a code, obviously $D = 2n + 1$, and the code allows one to correct n or fewer errors; here the whole space of $2^N = 2^{2n+1}$ points is decomposed into two Hamming spheres of radius n (each containing $2^{2n} = 2^{N-1}$ points). In addition to this, there is a non-trivial (and highly important) class of perfect codes formed by the Hamming codes with $N = 2^K - 1$, $M = 2^K - K - 1$. In fact, it has already been remarked on p. 315 that the Hamming inequality for the number of 'check signals' (which is equivalent to inequality (8)) becomes an equality for these codes; hence they are perfect. In the case of Hamming perfect codes, the whole space of $2^N = 2^{2^K-1}$ points is decomposed into 2^{2^K-K-1} Hamming spheres of radius 1, each of which contains 2^K points; here $D = 3$ and, consequently, all single errors can be corrected. But, if it is just assumed that $n > 1$ and $S > 2$, then we immediately encounter the foremost difficulty that for the existence of a perfect code the sum $1 + \binom{N}{1} + \dots + \binom{N}{n}$ by (8) must be equal to some integral power of 2 which in reality is seldom achieved. The American scientist M. J. E. Golay, in his search for perfect codes, noticed that

$$1 + \binom{23}{1} + \binom{23}{2} + \binom{23}{3} = 2048 = 2^{11},$$

and this suggested that in principle there may exist a perfect code with

$$N = 23 \text{ and } S = \frac{2^{23}}{2^{11}} = 2^{12} = 4096,$$

capable of correcting any combination of three or fewer errors. He indeed succeeded in finding such a code (since called the *Golay perfect binary code*). The code turns out to be a (23, 12)-cyclic parity-check code, defined by the generator polynomial

$$g(x) = x^{11} + x^9 + x^7 + x^6 + x^5 + x + 1,$$

or by the check polynomial

$$h(x) = \frac{x^{23} - 1}{g(x)} = x^{12} + x^{10} + x^7 + x^4 + x^3 + x^2 + x + 1$$

and having code distance $D = 7$.† Subsequent searches for new perfect codes have not been fruitful; except for those enumerated above, no other such codes have been found.†† This

The Golay code turned out to coincide also with the (nonprimitive) Bose-Chaudhuri-Hocqenghem code corresponding to the values $N = 23$ and $n = 2$ (i.e., correcting all single and double errors). However, the construction of this code by the method due to Bose *et al.*, allows one to assert only that for this $D \geq 5$ (expressly this also means that it is a double-error correcting code), whereas Golay claimed that in fact here $D = 7$.

††This apparently does not imply that there do not exist other sums of the form

$$1 + \binom{N}{1} + \dots + \binom{N}{n}$$

equal to a power of two. Thus, for instance, it is easy to verify that

$$1 + \binom{90}{1} + \binom{90}{2} = 2^{18}$$

but it can be nevertheless proved that there exists no perfect code with $N = 90$ and $n = 2$.

makes many scientists to infer that there are no perfect codes excepting those enumerated above. Attempts to prove rigorously that there are no new perfect codes succeeded in the early seventies due to Tietäväinen and Perko [217] in Finland and Zinoviev and Leontiev [220] in USSR. The initial results of these authors were related to perfect binary codes (i.e., codes of messages represented by sequences of two elementary signals). However, later Zinoviev and Leontiev [221] and Tietäväinen [216] independently obtained a complete solution of the problem of finding all perfect codes that employ p^k elementary signals, where p is an arbitrary prime and k is any positive integer. It turns out that such general perfect codes are quite rare, too.

Since perfect codes are so scarce, much attention has been devoted to the search for so-called *quasi-perfect* codes, slightly inferior to perfect codes but nevertheless sufficiently good. Quasi-perfect codes are defined as such codes that Hamming spheres of a certain fixed radius n with centres at the points corresponding to all possible code words fill out the entire space of 2^N points b , with the exception of only some $T < S \binom{N}{n+1}$ points (where S is the number of code words for the code) located at Hamming distance $n+1$ from at least one (but may be also from several) code word. If we agree in the case of quasi-perfect codes to decipher as $a^{(i)}$ all the received sequences b separated from the code word $a^{(i)}$ by not more than Hamming distance n and to decipher the sequence b separated by a distance $n+1$ from the code word closest to it as one (it is immaterial which) of the code words separated from b by a distance $n+1$, then inequality (9) also becomes an equality here. Hence, even for quasi-perfect codes used for transmission over a binary symmetric channel, *the mean probability of decoding error is less than for any other code with the same values of N and S* . At the same time, quasi-perfect codes exist in greater number than perfect codes (even though they are not many). Thus, for instance, the codes correcting all single errors in a block of $N \neq 2K-1$ digits and obtained by omitting a certain number of columns in a check matrix corresponding to the Hamming perfect code with $N = 2K-1$ quite often turn out to be quasi-perfect (see, for example [202], Chapter 5). The (primitive) double-error-correcting Bose-Chaudhuri-Hocquenghem codes with $N = 2K-1$, considered on pp. 334-336, are also all quasi-perfect (see, e.g., [202]); it is specifically on this basis that such codes were affirmed on p. 336 to be necessarily optimal. A series of other examples of quasi-perfect codes is described in Chapter 5 of [212]; we shall not further elaborate on this here.

Finally, let us describe the general construction of the binary Bose-Chaudhuri-Hocquenghem codes that have been mentioned repeatedly in this section. The basis of this construction is an ingenious description of the code generator polynomial by determining its roots, i.e., the solutions of the equation $g(x) = 0$. The main difficulty in determining the roots of $g(x)$ is easy to understand if we remember that the roots of ordinary polynomial with real coefficients may not be compulsorily real numbers, but may belong to a wider (i.e., the one containing a field of real numbers as its part) field of complex numbers. Quite similarly the roots of polynomial $g(x)$ with coefficients from a finite field may belong to the extension over a given field, i.e., to a new field containing the primary finite field as its part. In particular, when coefficients $g(x)$ are the elements of 2-arithmetic (a field $F_2 = \{0, 1\}$ with two elements 0 and 1), the roots of $g(x)$ may belong to a finite field F_{2^m} with 2^m distinct elements, where $m > 1$; see Appendix II. (As explained in Appendix II, the field F_{2^m} is nothing but the collection of all polynomials $a_0 + a_1\beta + \dots + a_{m-1}\beta^{m-1}$, where a_0, a_1, \dots, a_{m-1} are elements of $F_2 = \{0, 1\}$ and β is a root of irreducible polynomial $P_m(x)$ of degree m with all coefficients equal to 0 or 1. Another equivalent representation of F_{2^m} is given by $P_m(x)$ -arithmetic, i.e., the field of all the remainders after division of arbitrary polynomials by $P_m(x)$.)

Our problem is to find a generator polynomial $g(x)$ such that any pair of code polynomials $a(x)$ and $a_1(x)$ of the corresponding polynomial generator code has more than d distinct coefficients, where $d = 2n$ is a given integer (and n is the maximal number of errors to be correct-

ed in a code word; see p. 309). Let us first choose an integer r such that $2r - 1 > d$. Consider a finite field F_{2^r} of order 2^r constructed with the aid of an irreducible polynomial $P_r(x)$ of degree r with coefficients from $F_2 = \{0, 1\}$. Let α be a primitive element of F_{2^r} , i.e., all the consecutive powers $\alpha^1 = \alpha, \alpha^2, \alpha^3, \dots, \alpha^{2^r-1} = 1$ be distinct (see Appendix II, where it is also indicated that any $r + 1$ elements of the field F_{2^r} are linearly dependent, i.e., the sum of some of these $r + 1$ elements is equal to zero; recall that all coefficients λ_i of equation (7) of Appendix II are equal to 0 or 1 in our case). Let us consider the collection of elements

$$\alpha^0 = 1, \alpha^1 = \alpha, \alpha^2, \dots, \alpha^{n_1},$$

where n_1 is chosen subject to the condition that the elements $1, \alpha, \alpha^2, \dots, \alpha^{n_1}$ are linearly dependent but the elements $1, \alpha, \alpha^2, \dots, \alpha^{n_1-1}$ are not (it is clear that necessarily $n_1 < r + 1$). The corresponding linear dependence has the form

$$\alpha^{i_1^{(1)}} + \alpha^{i_2^{(1)}} + \dots + \alpha^{i_{k_1}^{(1)}} = 0, \quad (10)$$

where

$$0 \leq i_1^{(1)} < i_2^{(1)} < \dots < i_{k_1}^{(1)} = n_1 < r + 1$$

are certain integers. Here $x^{i_1^{(1)}} + x^{i_2^{(1)}} + \dots + x^{i_{k_1}^{(1)}} = 0$ is an equation of the lowest degree with the coefficients from F_2 having the root α . Accordingly, the polynomial

$$M_1(x) = x^{i_1^{(1)}} + x^{i_2^{(1)}} + \dots + x^{i_{k_1}^{(1)}}$$

may be called a *minimal polynomial* of the element α .

Consider now a sequence of consecutive powers of α^2 (i.e., $1, \alpha^2, \alpha^4, \alpha^6, \dots$) and let n_2 be the smallest number of the first terms of this sequence which are linearly dependent. Then

$$(\alpha^2)^{i_1^{(2)}} + (\alpha^2)^{i_2^{(2)}} + \dots + (\alpha^2)^{i_{k_2}^{(2)}} = 0, \quad (11)$$

where $i_{k_2}^{(2)} = n_2$. Here $M_2(x) = x^{i_1^{(2)}} + x^{i_2^{(2)}} + \dots + x^{i_{k_2}^{(2)}}$ is evidently the equation of the lowest degree having the root α^2 and hence $M_2(x)$ is the minimal polynomial of the element α^2 . (It will be shown later that $M_2(x)$ coincides with $M_1(x)$; however, this fact is immaterial here.) Similarly, we may consider the sequence $1, \alpha^4, \alpha^8, \dots$ and form an equation of the lowest degree having the root α^4 and the corresponding minimal polynomial $M_3(x)$. We continue to apply this procedure to $\alpha^4, \alpha^8, \dots, \alpha^d$.

Let us now consider the polynomial

$$g(x) = \text{lcm} [M_1(x), M_2(x), \dots, M_d(x)], \quad (12)$$

where lcm symbolizes the least common multiple of the polynomials in square brackets. It is possible to show that if $N = 2r - 1$, then $g(x)$ is just the desired generator polynomial. For this it is necessary to show that, if $g(x)$ is given by (12), then any two polynomials of the form (6) (see p. 329) of degree less than $2r - 1$ will have at least $d + 1$ distinct coefficients. Since

over the field F_2 (see Appendix II). Therefore,

$$(\alpha^3 + \alpha + 1)^2 = (\alpha^3)^2 + \alpha^2 + 1 = (\alpha^2)^3 + \alpha^2 + 1,$$

and hence $M_1(\alpha^2) = 0$, i.e., minimal polynomial $M_2(x)$ of α^2 coincides with $M_1(x)$. Quite similarly we can show that $[M_1'(\alpha^2)]^2 = M_1(\alpha^4)$ and hence $M_4(x) = M_1(x)$.†

Equations (15) imply, in particular, that $\alpha^9 = \alpha^2$ and that $\alpha^9 + \alpha^8 + 1 = 0$. Therefore, $(\alpha^3)^3 + (\alpha^3)^2 + 1 = 0$ and hence $M_3(x) = x^3 + x^2 + 1$ (it is easy to show that 1, α^3 and α^6 are linearly independent).

Now, we obtain

$$\begin{aligned} g(x) = \text{lcm}[M_1(x), M_2(x), M_3(x), M_4(x)] &= M_1(x)M_3(x) = (x^3 + x + 1)(x^3 + x^2 + 1) \\ &= x^6 + x^5 + x^4 + x^3 + x^2 + x + 1. \end{aligned}$$

Since $N = 2^3 - 1 = 7$ in our case, this generator polynomial corresponds to a very simple check polynomial

$$h(x) = \frac{x^7 - 1}{g(x)} = x - 1.$$

In the considered case $N = 7$ and the generator polynomial $g(x)$ is of the sixth degree. Hence we obtain a (not very advantageous !) double-error-correcting code with code words containing one information signal and six check signals. However, the general Varshamov-Gilbert inequality (****) on p. 327 shows that the number of check signals cannot be decreased here.

The simplest method to improve upon the proportion of information signals is to increase the value of r . Let us assume that $r = 4$ and hence $N = 2^r - 1 = 15$. The corresponding Hamming code correcting all single errors is a (15, 11)-code determined by the check matrix written on p. 325.†† In the case of double-error-correcting code we must consider a field F_{16} of $2^4 = 16$ elements determined by an irreducible fourth degree polynomial $Q(x)$ with coefficients from $F_2 = \{0, 1\}$. Let us choose $Q(x) = x^4 + x + 1$. It is easy to verify that this $Q(x)$ is irreducible over F_2 and a root α of $Q(x)$ is a primitive element of F_{16} represented by the collection of all quadrinoms $a_0 + a_1\alpha + a_2\alpha^2 + a_3\alpha^3$. In analogy to the above example, we can also show that here

$$M_1(x) = M_2(x) = M_4(x) = x^4 + x + 1,$$

$$M_3(x) = x^4 + x^3 + x^2 + x + 1.$$

Hence we obtain

$$g(x) = (x^4 + x + 1)(x^4 + x^3 + x^2 + x + 1).$$

This is the first generator polynomial written on p. 335.

†It is clear that this derivation is a general one : the minimal polynomials over a binary field $F_2 = \{0, 1\}$ always satisfy the relations : $M_1(x) = M_2(x) = M_4(x) = M_8(x) = \dots$, $M_3(x) = M_6(x) = M_{12}(x) = \dots$, $M_5(x) = M_{10}(x) = \dots$ and so on. These relations imply, in particular, that the inequality $K \leq dr$ (where $N = 2^r - 1$) implied by (12) can be replaced by a stronger inequality $K \leq dr/2$.

††The single-error-correcting Hamming codes are simultaneously the Bose-Chaudhuri-Hocquenghem codes corresponding to $d = 2$.

For a triple-error-correcting code (correcting all the single, double and triple errors) $d = 6$ and hence here also we can choose $r = 4$ (i.e., $N = 15$). We consider again the field F_{16} determined by the irreducible polynomial $Q(x) = x^4 + x + 1$. Let a root α of $Q(x)$ be again selected as a primitive element of F_{16} . As above, we obtain

$$M_1(x) = M_2(x) = M_4(x) = x^4 + x + 1, \quad M_3(x) = M_6(x) = x^4 + x^3 + x^2 + x + 1, \\ M_5(x) = x^2 + x + 1.$$

Hence in this case $g(x) = (x^2 + x + 1)(x^4 + x + 1)(x^4 + x^3 + x^2 + x + 1)$. This is the second generator polynomial written on p. 335.

The above construction of the Bose-Chaudhuri-Hocquenghem codes can be generalized quite easily to the case of a non-binary communication channel employing p^n elementary signals, where p is an arbitrary prime and n is an integer. For a study of this aspect, the reader is referred to books on coding and information theory mentioned on pp. 305-306; we shall not dwell upon it here.

APPENDIX I

Properties of convex functions

A function $y = f(x)$ is said to be *convex upward* (or, for short, simply *convex*) on an interval from $x = a$ to $x = b$ if in this interval every arc MN joining two points of the graph of the function lies *above* the corresponding *chord* MN † (Fig. 31). There are numerous examples of convex functions including the following: the logarithmic function $y = \log x$ in the entire domain, i.e., from 0 to ∞ ; the power function $y = -x^m$ in the same domain where $m > 1$; the exponential function $y = -a^x$ in the domain from $-\infty$ to $+\infty$; the function $y = -x \log x$ in the domain from 0 to ∞ , and the function

$$y = -x \log x - (1 - x) \log (1 - x)$$

in the domain from 0 to 1 (Fig. 32, $a - e$).

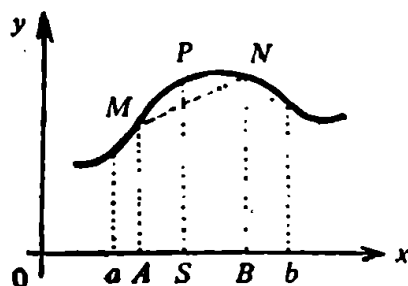


Fig. 31.

Theorem 1. If $y = f(x)$ is a convex function on an interval from a to b and x_1, x_2 are two values of the argument of this function within this interval (i.e., two arbitrary numbers such that $a \leq x_1 < x_2 \leq b$), then

$$\frac{f(x_1) + f(x_2)}{2} < f\left(\frac{x_1 + x_2}{2}\right). \quad (1)$$

Proof (Cf. also p. 48). Suppose that, in Fig. 31, $OA = x_1$, $OB = x_2$; in such a case $AM = f(x_1)$, $BN = f(x_2)$. Furthermore, if S is the centre of the segment AB , then $OS = (x_1 + x_2)/2$ and, consequently, $SP = f[(x_1 + x_2)/2]$. On the other hand, since the middle line SQ of the trapezium $ABNM$ is the mean of the base AM and BN , hence $SQ = [f(x_1) + f(x_2)]/2$. But, by the definition of

†In differential calculus it is shown that the following *convexity test* holds for a sufficiently wide class of functions (in particular, for all functions considered in this appendix): the function $y = f(x)$ is convex on the interval $a \leq x \leq b$ if its second derivative y'' is everywhere negative (i.e., $y'' < 0$) in this interval.

convex functions, the midpoint Q of chord MN lies below the point P of arc MN ; consequently,

$$\frac{f(x_1) + f(x_2)}{2} < f\left(\frac{x_1 + x_2}{2}\right),$$

giving the desired proof.†

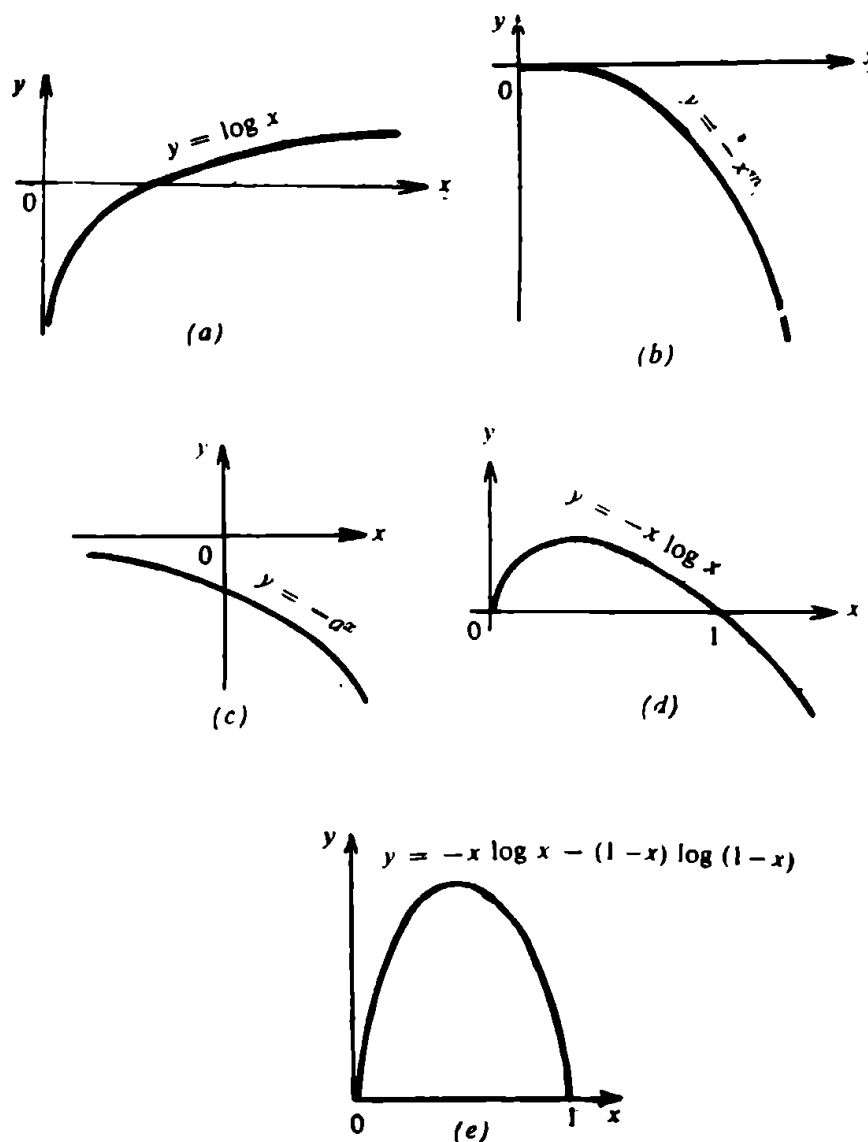


Fig. 32.

†Our proof is restricted to the case in which $f(x_1)$ and $f(x_2)$ have the same sign (in fact, only this case will be needed later by us). The reader may consider independently the case in which $f(x_1)$ and $f(x_2)$ have different signs (here the property of the middle line of the trapezium must be replaced by the following statement: *the segment of the middle line of a trapezium included between its diagonals is equal to the half of the difference of two trapezium bases*).

Examples†

(a) $y = \log x$. We have

$$\frac{\log x_1 + \log x_2}{2} < \log \frac{x_1 + x_2}{2},$$

i.e.,

$$\log \sqrt{x_1 x_2} < \log \frac{x_1 + x_2}{2},$$

or, finally,

$$\sqrt{x_1 x_2} < \frac{x_1 + x_2}{2}.$$

i.e., the geometric mean of two distinct positive numbers is less than their arithmetic mean.

(b) $y = -x^m$, $m > 1$. Here we obtain

$$-\frac{x_1^m + x_2^m}{2} < -\left(\frac{x_1 + x_2}{2}\right)^m,$$

or, in the different form,

$$\frac{x_1^m + x_2^m}{2} > \left(\frac{x_1 + x_2}{2}\right)^m, \left(\frac{x_1^m + x_2^m}{2}\right)^{1/m} > \frac{x_1 + x_2}{2}.$$

The expression

$$\left(\frac{a_1^m + a_2^m + \dots + a_k^m}{k}\right)^{1/m},$$

the m th root of the arithmetic mean of m th power numbers a_1, a_2, \dots, a_k , is called the *exponential mean of order m* of these k numbers (in particular, the expression

$$\sqrt{\left(\frac{a_1^2 + a_2^2 + \dots + a_k^2}{k}\right)},$$

†In the contents of this book we have substantially used only inequalities related to the convex functions $y = -x \log x$ and $y = \log x$ (as well as $y = -x \log x - (1-x) \log (1-x)$). Example (b) has here and hereafter only an illustrative value. (The theory of convex functions is, in fact, a rich source of curious inequalities, which permits us to multiply arbitrarily the number of examples.)

corresponding to the case $m = 2$, is called the *root-mean square* of the numbers a_1, a_2, \dots, a_k . Thus, the result obtained can be formally stated thus: *the exponential mean of order $m > 1$ of two distinct positive numbers is always greater than their arithmetic mean.*

(c) $y = -x \log x$. From Theorem 1 it follows that

$$-\frac{x_1 \log x_1 + x_2 \log x_2}{2} < -\frac{x_1 + x_2}{2} \log \frac{x_1 + x_2}{2},$$

or,

$$-\frac{1}{2}x_1 \log x_1 - \frac{1}{2}x_2 \log x_2 < -\frac{1}{2}(x_1 + x_2) \log \frac{x_1 + x_2}{2},$$

a conclusion that we have used twice (see pp. 49 and 65).

The inequality of Theorem 1 can be generalized as shown in the next theorem.

Theorem 2. *If the function $y = f(x)$ is convex in the interval from a to b , x_1 and x_2 are two arbitrary numbers in this interval (i.e., $a \leq x_1 < x_2 \leq b$) and p and q are some arbitrary positive numbers, whose sum is 1, then*

$$pf(x_1) + qf(x_2) < f(px_1 + qx_2). \quad (2)$$

For $p = q = \frac{1}{2}$, Theorem 2 reduces to Theorem 1.

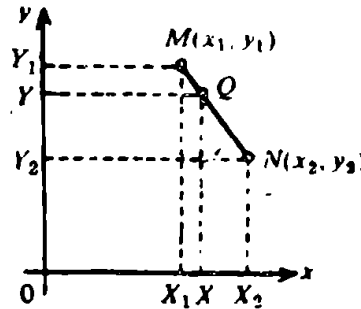


Fig. 33.

Proof. We first note that, if M and N are two points with coordinates (x_1, y_1) and (x_2, y_2) and Q is a point of the segment MN , dividing this segment in the ratio $MQ : QN = q : p$ (where $p + q = 1$), then the coordinates of the point Q are $px_1 + qx_2$ and $py_1 + qy_2$. Indeed, we denote by X_1, X_2 and X ; Y_1, Y_2 and Y the projections of points M, N and Q on the coordinate axes (Fig. 33). The points X and Y then divide the segments X_1X_2 and Y_1Y_2 in the ratio $q : p$. Hence†

$$OX = OX_1 + X_1X = x_1 + q(x_2 - x_1) = (1 - q)x_1 + qx_2 = px_1 + qx_2,$$

†Figure 33 depicts the case in which all four numbers x_1, x_2, y_1 and y_2 are positive (essentially this is the only case needed by us). The reader can independently examine other cases.

and

$$OY = OY_2 + Y_2Y = y_2 + p(y_1 - y_2) = (1 - p)y_2 + py_1 = py_1 + qy_2.$$

We now consider again the graph of our convex function $y = f(x)$ (Fig. 34),

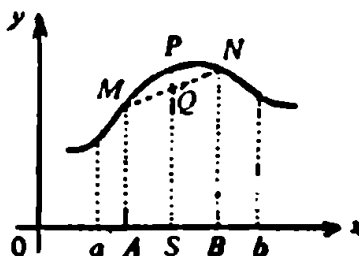


Fig. 34.

and let $OA = x_1$, $OB = x_2$, $AM = f(x_1)$, $BN = f(x_2)$. By what has been proved above, the coordinates of the point Q , dividing the segment MN in the ratio

$$MQ : QN = q : p$$

are $px_1 + qx_2$ and $pf(x_1) + qf(x_2)$. Thus, in Fig. 34, $SQ = pf(x_1) + qf(x_2)$ and $SP = f(px_1 + qx_2)$ (because $OS = px_1 + qx_2$). But, because of the convexity of $y = f(x)$, the point Q is located *below* the point P ; hence,

$$pf(x_1) + qf(x_2) < f(px_1 + qx_2),$$

giving the desired proof.†

Examples

(a) $y = \log x$. In this case, inequality (2) yields

$$p \log x_1 + q \log x_2 < \log (px_1 + qx_2).$$

Hence it follows that

$$x_1^p x_2^q < px_1 + qx_2, \quad p + q = 1.$$

(b) $y = -x^m$, $m > 1$. We have

†It is trivial to see that the coordinates of *each* point of the segment MN can be represented in the form $(px_1 + qx_2, py_1 + qy_2)$, with $p > 0$, $q > 0$, $p + q = 1$. Thus, inequality (2) says that the entire chord MN lies below the curve $y = f(x)$, i.e., it is equivalent to the definition of a convex function.

$$-px_1^m - qx_2^m < -(px_1 + qx_2)^m,$$

or

$$px_1^m + qx_2^m > (px_1 + qx_2)^m, \quad p + q = 1.$$

(c) $y = -x \log x$. Here we obtain

$$-px_1 \log px_1 - qx_2 \log qx_2 < -(px_1 + qx_2) \log (px_1 + qx_2), \quad p + q = 1.$$

Theorem 1 can also be generalized in another direction.

Theorem 3. *If $y = f(x)$ is a convex function in the interval from a to b and x_1, x_2, \dots, x_k are any k values of the argument of the function in this interval, none of which is equal to any of the others, then*

$$\frac{f(x_1) + f(x_2) + \dots + f(x_k)}{k} < f\left(\frac{x_1 + x_2 + \dots + x_k}{k}\right) \quad (3)$$

(a particular case of Jensen's inequality).

For $k = 2$, Theorem 3 reduces to Theorem 1.

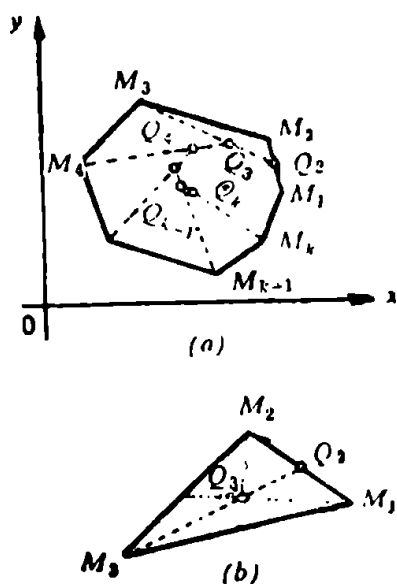


Fig. 35.

Proof. To start with, we define a concept frequently encountered in geometric and analytic problems. Suppose that $M_1M_2M_3 \dots M_k$ is an arbitrary k -gon (Fig. 35a). Let us also assume that Q_2 is the midpoint of the side M_1M_2 of this k -gon ($M_1Q_2 : Q_2M_2 = \frac{1}{2} : \frac{1}{2}$); Q_3, Q_4, \dots, Q_k are the points that divide

the segments $M_3Q_2, M_4Q_3, \dots, M_kQ_{k-1}$, respectively, in the ratios $2:1$ (i.e., $M_3Q_3:Q_3Q_2 = \frac{2}{3}:\frac{1}{3}$), $3:1$ (i.e., $M_4Q_4:Q_4Q_3 = \frac{3}{4}:\frac{1}{4}$), \dots , $(k-1):1$ (i.e.,

$$M_kQ_k:Q_kQ_{k-1} = \frac{k-1}{k}:\frac{1}{k}.$$

The point Q_k is called the *centroid* (or the *centre of gravity*) of k -gon $M_1M_2\dots M_k$. In the case of the triangle $M_1M_2M_3$ (Fig. 35b) the centroid Q_3 is the *point of intersection of its medians*: indeed, in this case Q_2 is the midpoint of the side M_1M_2 , the segment M_3Q_2 is a median and the point Q_3 , dividing this segment in the ratio $M_3Q_3:Q_3Q_2 = 2:1$, is a point of intersection of the medians of the triangle.

Let us now show that, if the coordinates of the vertices M_1, M_2, \dots, M_k of a k -gon are $(x_1, y_1), (x_2, y_2), \dots, (x_k, y_k)$, then the coordinates of the centroid Q_k are $(x_1 + x_2 + \dots + x_k)/k$ and $(y_1 + y_2 + \dots + y_k)/k$.† Indeed, by the propositions deduced in the beginning of the proof of Theorem 2, the points Q_2, Q_3, Q_4, \dots , and, finally, Q_k have the following coordinates:

$$Q_2\left(\frac{x_1 + x_2}{2}, \frac{y_1 + y_2}{2}\right),$$

$$Q_3\left(\frac{2}{3}\frac{x_1 + x_2}{2} + \frac{1}{3}x_3, \frac{2}{3}\frac{y_1 + y_2}{2} + \frac{1}{3}y_3\right),$$

or, what is the same,

$$\left(\frac{x_1 + x_2 + x_3}{3}, \frac{y_1 + y_2 + y_3}{3}\right),$$

$$Q_4\left(\frac{3}{4}\frac{x_1 + x_2 + x_3}{3} + \frac{1}{4}x_4, \frac{3}{4}\frac{y_1 + y_2 + y_3}{3} + \frac{1}{4}y_4\right),$$

or

$$\left(\frac{x_1 + x_2 + x_3 + x_4}{4}, \frac{y_1 + y_2 + y_3 + y_4}{4}\right),$$

\dots

$$Q_k\left(\frac{(k-1)}{k}\frac{x_1 + x_2 + \dots + x_{k-1}}{k-1} + \frac{1}{k}x_k, \frac{(k-1)}{k}\frac{y_1 + y_2 + \dots + y_{k-1}}{k-1} + \frac{1}{k}y_k\right),$$

†Hence, it follows in particular that the centroid of a k -gon is completely determined by this k -gon and does not depend on the order of enumeration of its vertices (as can be believed from the definition of a centroid). In the case of a triangle this last circumstance also stems from the fact that the centroid of a triangle is the point of intersection of the medians.

or

$$\left(\frac{x_1 + x_2 + \dots + x_{k-1} + x_k}{k}, \frac{y_1 + y_2 + \dots + y_{k-1} + y_k}{k} \right).$$

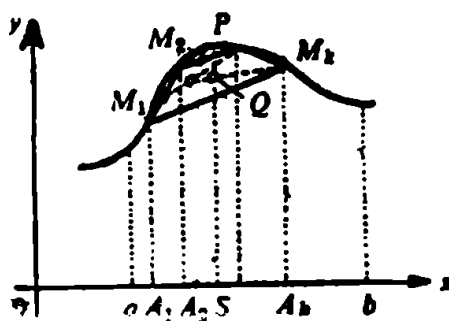


Fig. 36.

We now recall our convex function $y = f(x)$. Suppose that M_1, M_2, \dots, M_k are k successive points of the graph of this function within the considered interval (Fig. 36). Because of the convexity of the function, the k -gon $M_1M_2 \dots M_k$ is convex and lies wholly *below* the curve $y = f(x)$. If the abscissas of the points M_1, M_2, \dots, M_k are x_1, x_2, \dots, x_k , then their ordinates are obviously $f(x_1), f(x_2), \dots, f(x_k)$. Hence, the coordinates of the centroid Q of k -gon $M_1M_2 \dots M_k$ are given by

$$\frac{x_1 + x_2 + \dots + x_k}{k} \quad \text{and} \quad \frac{f(x_1) + f(x_2) + \dots + f(x_k)}{k},$$

and, consequently,

$$OS = \frac{x_1 + x_2 + \dots + x_k}{k}, \quad SQ = \frac{f(x_1) + f(x_2) + \dots + f(x_k)}{k},$$

and

$$SP = f\left(\frac{x_1 + x_2 + \dots + x_k}{k}\right)$$

(see Fig. 36). However, the centroid of a convex k -gon lies always *interior* to the k -gon (this is implied by the very definition of a centroid). Consequently, the point Q lies *below* the point P and, hence,

$$\frac{f(x_1) + f(x_2) + \dots + f(x_k)}{k} < f\left(\frac{x_1 + x_2 + \dots + x_k}{k}\right),$$

giving the required proof.

This reasoning is conserved also in the case in which some of the points M_1, M_2, \dots, M_k (but not all!) coincide (some of the numbers x_1, x_2, \dots, x_k are equal among themselves). The k -gon $M_1 M_2 \dots M_k$ is here obviously expressed as a polygon with a smaller number of vertices.

Examples

(a) $y = \log x$. From Theorem 3 it follows that

$$\frac{\log x_1 + \log x_2 + \dots + \log x_k}{k} < \log \frac{x_1 + x_2 + \dots + x_k}{k},$$

or

$$\sqrt[k]{x_1 x_2 \dots x_k} < \frac{x_1 + x_2 + \dots + x_k}{k}.$$

We see that *the geometric mean of k positive numbers, at least two of which are distinct, is less than their arithmetic mean* (the theorem on geometric and arithmetic means).

(b) $y = -x^m, m > 1$. In such a case we obtain

$$-\frac{x_1^m + x_2^m + \dots + x_k^m}{k} < -\left(\frac{x_1 + x_2 + \dots + x_k}{k}\right)^m,$$

or

$$\left(\frac{x_1^m + x_2^m + \dots + x_k^m}{k}\right)^{1/m} > \frac{x_1 + x_2 + \dots + x_k}{k}.$$

This shows that *the exponential mean of order $m > 1$ of any k positive numbers, at least two of which are distinct, is greater than their arithmetic mean*.

(c) $y = -x \log x$. In this case, Theorem 3 yields

$$\begin{aligned} & -\frac{x_1 \log x_1 + x_2 \log x_2 + \dots + x_k \log x_k}{k} \\ & < -\frac{x_1 + x_2 + \dots + x_k}{k} \log \left(\frac{x_1 + x_2 + \dots + x_k}{k} \right). \quad (4) \end{aligned}$$

Finally, we shall prove one more theorem, which is an extension of both Theorems 2 and 3,

$$Q_2 \left(\frac{p_1 x_1 + p_2 x_2}{p_1 + p_2}, \frac{p_1 f(x_1) + p_2 f(x_2)}{p_1 + p_2} \right),$$

$$Q_3 \left(\frac{p_1 + p_2}{p_1 + p_2 + p_3} \frac{p_1 x_1 + p_2 x_2}{p_1 + p_2} + \frac{p_3}{p_1 + p_2 + p_3} x_3, \right.$$

$$\left. \frac{p_1 + p_2}{p_1 + p_2 + p_3} \frac{p_1 f(x_1) + p_2 f(x_2)}{p_1 + p_2} + \frac{p_3}{p_1 + p_2 + p_3} f(x_3) \right),$$

or

$$\left(\frac{p_1 x_1 + p_2 x_2 + p_3 x_3}{p_1 + p_2 + p_3}, \frac{p_1 f(x_1) + p_2 f(x_2) + p_3 f(x_3)}{p_1 + p_2 + p_3} \right),$$

$$Q_4 \left(\frac{p_1 + p_2 + p_3}{p_1 + p_2 + p_3 + p_4} \frac{p_1 x_1 + p_2 x_2 + p_3 x_3}{p_1 + p_2 + p_3} + \frac{p_4}{p_1 + p_2 + p_3 + p_4} x_4, \right.$$

$$\left. \frac{p_1 + p_2 + p_3}{p_1 + p_2 + p_3 + p_4} \frac{p_1 f(x_1) + p_2 f(x_2) + p_3 f(x_3)}{p_1 + p_2 + p_3} + \frac{p_4}{p_1 + p_2 + p_3 + p_4} f(x_4) \right),$$

or

$$\left(\frac{p_1 x_1 + p_2 x_2 + p_3 x_3 + p_4 x_4}{p_1 + p_2 + p_3 + p_4}, \frac{p_1 f(x_1) + p_2 f(x_2) + p_3 f(x_3) + p_4 f(x_4)}{p_1 + p_2 + p_3 + p_4} \right),$$

$$\dots \dots \dots$$

$$Q \left(\frac{p_1 x_1 + p_2 x_2 + \dots + p_{k-1} x_{k-1} + p_k x_k}{p_1 + p_2 + \dots + p_{k-1} + p_k}, \right.$$

$$\left. \frac{p_1 f(x_1) + p_2 f(x_2) + \dots + p_{k-1} f(x_{k-1}) + p_k f(x_k)}{p_1 + p_2 + \dots + p_{k-1} + p_k} \right),$$

or, differently,

$$(p_1 x_1 + p_2 x_2 + \dots + p_k x_k, p_1 f(x_1) + p_2 f(x_2) + \dots + p_k f(x_k))$$

(since $p_1 + p_2 + \dots + p_k = 1$).

Thus, in Fig. 37,

$$SQ = p_1 f(x_1) + p_2 f(x_2) + \dots + p_k f(x_k),$$

$$OS = p_1 x_1 + p_2 x_2 + \dots + p_k x_k,$$

$$SP = f(p_1 x_1 + p_2 x_2 + \dots + p_k x_k).$$

But since the point Q lies *below* the point P (because the entire k -gon $M_1 M_2 \dots M_k$ lies below the curve $y = f(x)$, and Q is an *interior* point of this k -gon), we have

$$p_1 f(x_1) + p_2 f(x_2) + \dots + p_k f(x_k) < f(p_1 x_1 + p_2 x_2 + \dots + p_k x_k),$$

giving the required proof.†

Examples

(a) $y = \log x$. We have

$$p_1 \log x_1 + p_2 \log x_2 + \dots + p_k \log x_k < \log (p_1 x_1 + p_2 x_2 + \dots + p_k x_k),$$

implying that

$$x_1^{p_1} x_2^{p_2} \dots x_k^{p_k} < p_1 x_1 + p_2 x_2 + \dots + p_k x_k,$$

where

$$p_1 + p_2 + \dots + p_k = 1$$

(the generalized theorem on the geometric and arithmetic means).

(b) $y = -x^m$, $m > 1$. We have

$$-p_1 x_1^m - p_2 x_2^m - \dots - p_k x_k^m < -(p_1 x_1 + p_2 x_2 + \dots + p_k x_k)^m,$$

or

$$p_1 x_1^m + p_2 x_2^m + \dots + p_k x_k^m > (p_1 x_1 + p_2 x_2 + \dots + p_k x_k)^m,$$

where

$$p_1 + p_2 + \dots + p_k = 1.$$

(c) $y = -x \log x$. Theorem 4 yields

$$\begin{aligned} & -p_1 x_1 \log x_1 - p_2 x_2 \log x_2 - \dots - p_k x_k \log x_k \\ & < -(p_1 x_1 + p_2 x_2 + \dots + p_k x_k) \log (p_1 x_1 + p_2 x_2 + \dots + p_k x_k), \end{aligned}$$

where

$$p_1 + p_2 + \dots + p_k = 1. \quad (6)$$

The derivation of inequalities (4) on p. 355 and (6) is the basic aim of this appendix. From inequality (4) it is immediate that the entropy of an experiment

†It is trivial to see that the coordinates of every interior point of the k -gon $M_1 M_2 \dots M_k$ can be represented in the form

$$(p_1 x_1 + p_2 x_2 + \dots + p_k x_k, p_1 f(x_1) + p_2 f(x_2) + \dots + p_k f(x_k)),$$

where $p_1 > 0, p_2 > 0, \dots, p_k > 0$, and $p_1 + p_2 + \dots + p_k = 1$. Thus, inequality (5) expresses the situation that a polygon inscribed in the graph of a convex function wholly lies below this graph.

α having k outcomes does not exceed the entropy $\log k$ of an experiment α_0 with k outcomes of *equal probability*; also, $H(\alpha) = \log k$ if and only if all outcomes of α are equally probable, i.e., if α is not different from α_0 . Actually, we multiply both sides of inequality (4) by k and then substitute in this

$$x_1 = p(A_1), x_2 = p(A_2), \dots, x_k = p(A_k),$$

where A_1, A_2, \dots, A_k are outcomes of α (so that $p(A_1) + p(A_2) + \dots + p(A_k) = 1$; the probabilities $p(A_1), p(A_2), \dots, p(A_k)$ are not all equal among themselves). In such a case, we have

$$\begin{aligned} & -p(A_1) \log p(A_1) - p(A_2) \log p(A_2) - \dots - p(A_k) \log p(A_k) \\ & < -[p(A_1) + p(A_2) + \dots + p(A_k)] \\ & \quad \times \log \frac{p(A_1) + p(A_2) + \dots + p(A_k)}{k} \\ & = -1 \times \log \frac{1}{k} = \log k, \end{aligned}$$

or

$$H(\alpha) < H(\alpha_0).$$

Inequality (6) can be used to prove that the conditional entropy $H_\alpha(\beta)$ of β given α does not exceed the unconditional entropy $H(\beta)$ of β . In fact, let us put in inequality (6)

$$p_1 = p(A_1), p_2 = p(A_2), \dots, p_k = p(A_k),$$

$$x_1 = p_{A_1}(B_1), x_2 = p_{A_2}(B_1), \dots, x_k = p_{A_k}(B_1)$$

(where A_1, A_2, \dots, A_k and B_1, B_2, \dots, B_l are outcomes of α and β ; $p(A_1) + p(A_2) + \dots + p(A_k) = 1$). Then, we obtain

$$\begin{aligned} & -p(A_1)p_{A_1}(B_1) \log p_{A_1}(B_1) - p(A_2)p_{A_2}(B_1) \log p_{A_2}(B_1) - \dots \\ & \quad - p(A_k)p_{A_k}(B_1) \log p_{A_k}(B_1) \\ & < -[p(A_1)p_{A_1}(B_1) + p(A_2)p_{A_2}(B_1) + \dots + p(A_k)p_{A_k}(B_1)] \\ & \quad \times \log [p(A_1)p_{A_1}(B_1) + p(A_2)p_{A_2}(B_1) + \dots + p(A_k)p_{A_k}(B_1)]. \end{aligned}$$

Since by the equation of total probability (see p. 23)

$$p(A_1)p_{A_1}(B_1) + p(A_2)p_{A_2}(B_1) + \dots + p(A_k)p_{A_k}(B_1) = p(B_1),$$

hence

$$-p_1 \log p_1 - p_2 \log p_2 - \dots - p_k \log p_k < -p_1 \log q_1 - p_2 \log q_2 - \dots - p_k \log q_k,$$

i.e., we arrive at inequality (*) on p. 134.

Finally, let us consider the inequality

$$H_{\alpha\gamma}(\beta) < H_{\gamma}(\beta),$$

which generalizes inequality $H_{\alpha}(\beta) \leq H(\beta)$, and has been mentioned at the end of Sec. 3 of Chap. 2. (This inequality turns into $H_{\alpha}(\beta) \leq H(\beta)$ if it is assumed that experiment γ has a single outcome realized with probability 1.) It is easy to deduce the considered inequality from inequality $H_{\alpha}(\beta) \leq H(\beta)$. Indeed, we denote by C_1, C_2, \dots, C_m the outcomes of an experiment γ ; suppose $\alpha^{(1)}$ and $\beta^{(1)}$ to be experiments with outcomes

$$A_1^{(1)}, A_2^{(1)}, \dots, A_k^{(1)} \text{ and } B_1^{(1)}, B_2^{(1)}, \dots, B_l^{(1)}$$

having probabilities

$$p(A_1^{(1)}) = p_{C_1}(A_1), p(A_2^{(1)}) = p_{C_1}(A_2), \dots, p(A_k^{(1)}) = p_{C_1}(A_k),$$

and

$$p(B_1^{(1)}) = p_{C_1}(B_1), p(B_2^{(1)}) = p_{C_1}(B_2), \dots, p(B_l^{(1)}) = p_{C_1}(B_l),$$

respectively. By what has been proved above, we have

$$H_{\alpha^{(1)}}(\beta^{(1)}) < H(\beta^{(1)}).$$

But

$$\begin{aligned} H(\beta^{(1)}) &= -p(B_1^{(1)}) \log p(B_1^{(1)}) - p(B_2^{(1)}) \log p(B_2^{(1)}) - \dots - p(B_l^{(1)}) \log p(B_l^{(1)}) \\ &= -p_{C_1}(B_1) \log p_{C_1}(B_1) - p_{C_1}(B_2) \log p_{C_1}(B_2) - \dots - p_{C_1}(B_l) \log p_{C_1}(B_l) = H_{C_1}(\beta) \end{aligned}$$

and

$$H_{\alpha^{(1)}}(\beta^{(1)}) = p(A_1^{(1)})H_{A_1^{(1)}}(\beta^{(1)}) + p(A_2^{(1)})H_{A_2^{(1)}}(\beta^{(1)}) + \dots + p(A_k^{(1)})H_{A_k^{(1)}}(\beta^{(1)}),$$

where

$$\begin{aligned} H_{A_1^{(1)}}(\beta^{(1)}) &= -p_{A_1^{(1)}}(B_1^{(1)}) \log p_{A_1^{(1)}}(B_1^{(1)}) - p_{A_1^{(1)}}(B_2^{(1)}) \log p_{A_1^{(1)}}(B_2^{(1)}) - \dots \\ &\quad \dots - p_{A_1^{(1)}}(B_l^{(1)}) \log p_{A_1^{(1)}}(B_l^{(1)}), \end{aligned}$$

APPENDIX II

Some algebraic concepts

The main subject of study in algebra is some *algebraic systems*, i.e., sets of elements, for which there are defined some *algebraic operations*, similar to the well-known arithmetic operations of addition and multiplication of numbers. Moreover, the nature of the elements of such a system and the concrete meaning of the operations under consideration are usually not specified, so that one and the same algebraic scheme can describe many diverse examples. On the contrary, the properties of algebraic operations are described explicitly, and this description forms the *definition* of the corresponding system.

1. The first algebraic concept, extensively used in many branches of mathematics, is the concept of a (*commutative*) *group*.

A set G of elements a, b, c, \dots is called a (commutative) group if on this set an operation \circ is defined, assigning to each pair of elements a and b of our set a unique third element denoted by the symbol $a \circ b$, and for which the following properties hold :

G1 : *The operation \circ is commutative :†*

$$a \circ b = b \circ a \text{ for every } a \text{ and } b \text{ in } G;$$

G2 : *The operation \circ is associative :*

$$(a \circ b) \circ c = a \circ (b \circ c) \text{ for every } a, b \text{ and } c \text{ in } G;$$

G3 : *In the set G there is an identity element e such that*

$$a \circ e = a \text{ for every } a \text{ in } G;$$

G4 : *For each a in G there is a symmetric element a^* such that*

$$a \circ a^* = e.$$

The group operation \circ is sometimes denoted by the symbol $+$ (additive group notation). In such a case, the element $a + b$ is called the *sum* of elements a and b ; an identity element e such that

$$a + e = a \text{ for every } a$$

†In algebra *non-commutative* groups are also often considered, for which the property G1 does not hold. However, since only commutative groups are encountered in this book, we have agreed, in departure from the convention, to include G1 in the definition of a group.

is called the *null* or *zero element* or simply the *null* or *zero* of the group and usually denoted by 0; a symmetric element a^* such that

$$a + a^* = 0$$

is called the *negative* of a and is written $-a$. The result $a \circ b$ of applying the group operation \circ to the elements a and b may also be written as $a \times b$ or ab (multiplicative group notation). In such a case, $ae = a$ for every a , and hence e is called the *unit element* or the *unit* of a group and sometimes denoted by 1; furthermore, $aa^* = 1$ and hence a^* is called here the *inverse* of a and written a^{-1} . We shall hereafter always denote a group operation by the symbol $+$; for this, we denote by $a - b$ an element x (the *difference* of elements a and b) such that $x + b = a$ (it is trivial to see that such element x always exists: it is equal to $a + (-b)$).

Examples

A. *A set of integers (or rational numbers, or real numbers) forms a group with respect to addition.* In other words, the corresponding set, where (ordinary) addition is taken as the group operation, forms a group with the null element 0 and the negative $-a$ of a .

B. We agree to take the *multiplication* of numbers as a group operation (which we now denote by the symbol $\ll + \gg$, in order to emphasize that this *is not* ordinary addition). For this, a set of integers, however, does not form a group, because here G4 is obviously not satisfied: in fact, an integer a^* such that $a \ll + \gg a^* = aa^* = 1$ exists if and only if $a = 1$ or $a = -1$. Similarly, a set of all rational numbers as well does not form a group with respect to multiplication because here G4 is violated for $a = 0$. However, *a set of all nonzero (or positive) rational numbers (or real numbers) forms a multiplicative group.*

C. We consider again a set of integers and define on this set an operation of addition of numbers. We now choose any positive integer q and agree to *replace every integer A by the remainder after the division of A by q* . Thus, say, if $q = 10$, then we agree to leave from each positive integer A only its last digit a (this also is the remainder obtained from the division of A by 10). A set of all possible *remainders obtained from the division of all integers by q* (formed of q numbers $0, 1, 2, \dots, q - 1$) is called a *q -arithmetic*; the sum of the elements a and b of a q -arithmetic is the remainder obtained from the division of the usual sum $a + b$ by q ($= a + b$ if $a + b < q$). Tables of addition in 2-arithmetic, 5-arithmetic and 6-arithmetic might look like the accompanying Tables 1, 2 and 3,

TABLE 1

+	0	1
0	0	1
1	1	0

TABLE 2

+	0	1	2	3	4
0	0	1	2	3	4
1	1	2	3	4	0
2	2	3	4	0	1
3	3	4	0	1	2
4	4	0	1	2	3

TABLE 3

+	0	1	2	3	4	5
0	0	1	2	3	4	5
1	1	2	3	4	5	0
2	2	3	4	5	0	1
3	3	4	5	0	1	2
4	4	5	0	1	2	3
5	5	0	1	2	3	4

It is easy to see that a q -arithmetic defined with respect to addition is itself a group of q elements (or as is said, is a group of order q). The null element of this group is 0 and the negative of $a \neq 0$ is the number $q - a$ (because the sum $a + (q - a)$ when divided by q gives the remainder 0). For 2-arithmetic, the negative of every number a (i.e., both for $a = 0$ and $a = 1$) is obviously a itself: here $-a = a$ always.

D. Suppose that G is an arbitrary group, say, a group of integers with respect to addition or a group of additions of numbers in a q -arithmetic. We now consider an arbitrary rectangular array of m rows and n columns, or an $(m \times n)$ -matrix

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}$$

composed of the elements of G , which we shall hereafter call *numbers*. It is clear that if we agree to add the matrices elementwise (i.e., consider that a number appearing at some place in the matrix-sum is equal to the sum of the numbers occurring at the same places in the matrix-summands), then we arrive at an additive group of $(m \times n)$ -matrices; the null element of this group is the zero matrix O , which has all zeros.

$(1 \times n)$ -matrices are also called *vectors* (or, *row vectors*); similarly, $(m \times 1)$ -matrices are called *column vectors*. Obviously, vectors with a fixed number of elements in row (or, column) also admit addition with each other; if the elements of a vector belong to some group ('group of numbers'), then the set of all vectors also forms a group with respect to addition. Vectors are mostly denoted by small bold-face Latin letters; the 'null vectors' (i.e., a row or a column composed of 0s) is sometimes denoted by a boldface 0.

If a group G of 'numbers' is infinite, then a corresponding group of $(m \times n)$ -matrices (in particular, of vectors) is also infinite. If, however, G is of finite order q , then a group of $(m \times n)$ -matrices is of order q^{mn} ; in fact, a matrix has

mn elements, in place of each of which we can substitute any of the q elements of G . Similarly, a group of row vectors of n elements and a group of column vectors of m elements are, respectively, of finite order q^n and q^m if G is of order q .

E. Consider an arbitrary *polynomial*

$$f(x) = a_0 + a_1x + a_2x^2 + \dots + a_{n-1}x^{n-1},$$

whose coefficients a_0, a_1, \dots, a_{n-1} are the elements of an arbitrarily chosen group G . If $g(x)$ is another polynomial

$$g(x) = b_0 + b_1x + b_2x^2 + \dots + b_{n-1}x^{n-1}$$

(we assume that $f(x)$ and $g(x)$ are of the same degree because otherwise there can always be added to the one of lower degree some 'leading' terms with the coefficient 0 under them), then the *sum* of polynomials can be defined by

$$f(x) + g(x) = (a_0 + b_0) + (a_1 + b_1)x + (a_2 + b_2)x^2 + \dots + (a_{n-1} + b_{n-1})x^{n-1}.$$

It is easy to see that the *polynomials* with addition so defined *form a group*. Obviously, this group is always infinite, because the degree of the polynomials can be arbitrarily large. The role of null element of this group is played by the 'null' polynomial 0, all of whose coefficients are 0; the negative of $f(x)$ is the polynomial $-f(x)$, all of whose coefficients are the negative of the coefficients of $f(x)$.

If we confine ourselves to *polynomials of degree less than n* , where n is some fixed number, then also we obtain a *group*; as is easy to see, it differs from the group of vectors

$$f = (a_0, a_1, a_2, \dots, a_{n-1})$$

only in the form of writing the elements of a group. This group is finite, if the group G is finite; if G is of order q , then the order of a group of polynomials of degree $< n$ is q^n . Thus, say, there are in all $2^2 = 4$ polynomials of degree < 2 with coefficients from a 2-arithmetical: 0, 1, x and $x + 1$; a 'table of addition' of these polynomials looks like the accompanying Table 4.

TABLE 4

+	0	1	x	$x + 1$
0	0	1	x	$x + 1$
1	1	0	$x + 1$	x
x	x	$x + 1$	0	1
$x + 1$	$x + 1$	x	1	0

Suppose now that G is an arbitrary group and that H is a subset of elements of G . If a set H of elements of a group is such that

SG1 : if a, b belong to H , then $a + b$ also belongs to H ;

SG2 : if a belongs to H , then $-a$ also belongs to H ;

SG3 : the null element 0 of G belongs to H ,

then H itself forms a group with respect to addition defined on G . In such a case, we say that H is a *subgroup* of G .

It is easy to see that a subgroup can also be defined as a set H of elements of a group satisfying the unique requirement: if a and b belong to H , then $a - b$ also belongs to H . In fact, then, evidently, 0 belongs to H , since $0 = a - a$, where a is any element of H . Moreover, if a belongs to H , then $-a$ also belongs to H , since $-a = 0 - a$; also if a and b belong to H , then $a + b = a - (-b)$ belongs to H .

In particular, if G is an additive group of integers, then a collection H of all integers that are multiples of a fixed integer l forms a subgroup of G . In exactly the same way, if G is an additive group of numbers in a q -arithmetic and $q = kl$ is a composite number, then a collection H of all numbers belonging to G that are divisible by l (i.e., the numbers $l, 2l, 3l, \dots, (k-1)l$) forms a subgroup of G (which differs immaterially, as is easy to understand, from an additive group of numbers in a k -arithmetic).

A subgroup of an additive group of $(m \times n)$ -matrices is, for example, a group of all matrices, all of whose rows, except the first one, contain only the zeros (this subgroup is obviously equivalent to and only written differently from an additive group of row vectors), and also a group of matrices, all of whose elements are 0 except a fixed one, say, the element a_{11} , appearing in the left top corner (this subgroup reduces to the group G , because each of its elements is given by the single number a_{11}). Furthermore, if G is a group of $(m \times n)$ -matrices A with elements from a 2-arithmetic, then in order to ensure that some subset of it is a subgroup it suffices only to verify that SG1 is satisfied (because in a 2-arithmetic every number is the inverse of itself and hence here $A + A = O$ for each matrix A and, consequently, $-A = A$).

A subgroup of a group of all polynomials is a group of polynomials of degree $< n$. For this, the latter group, a set of all polynomials of degree $< k$, where $k < n$, and a set of all polynomials that vanish for $x = 0$ (i.e., polynomials that have zero 'free term' a_0) both form a subgroup.

If H is a subgroup of G , then a set of all elements of the form $a + h$, where a is a fixed element of G and h runs through all elements of H , is called a *coset* of H in G and is written $a + H$. It is easy to show that any two cosets of H in G are either disjoint (i.e., they do not contain any common element), or are exactly the same. In fact, if the cosets $a + H$ and $b + H$ have a common

element, then $a + h_1 = b + h_2$, where both h_1 and h_2 belong to H , and hence $a - b = h_2 - h_1$, i.e., $a - b = h$ also belongs to H . Therefore, the coset $a + H$ can be represented as $b + h + H = B + (h + H)$. But if h belongs to H , then $h + H = H$, since any element h_1 of H can be represented as an element $h + (h_1 - h)$ of $h + H$, and any element $h + h_1$ of $h + H$ belongs to H . This completes the proof of the italicized assertion.

We see that a subgroup H determines the partition of a group G into disjoint cosets of H . If H contains a finite number n of elements, then any coset will contain n elements, too. Let us now assume that G is of finite order N (i.e., it contains N elements). Since all these elements must form a finite number of cosets, we obtain the following Lagrange's theorem.

LAGRANGE'S THEOREM. *If G is a finite group of order N and H a subgroup of G of order n , then $N = nk$, where k is an integer, i.e., n is a divisor of N . The integer k is, of course, equal to the number of cosets of H in G and called the index of H in G .*

Let G be a finite group and a an arbitrary element of G . Consider the sequence of sums

$$a = 0 + a, a + a = 2a, 2a + a = 3a, 3a + a = 4a, \dots$$

All these sums cannot be distinct since the number of distinct elements of G is finite. Moreover, if $ia = ja$, where (say) $j > i$, then $ja - ia = (j - i)a = 0$. The smallest integer n satisfying the relation $na = 0$ is called the *order* of an element a . It is easy to see that the elements $1a = a, 2a, 3a, \dots, (n - 1)a, na = 0$ form a subgroup of G which we call a *cyclic subgroup of G generated by a* . (If, however, the group G coincides with one of its cyclic subgroups, then G is called a *cyclic group*.) The order of the cyclic subgroup clearly coincides with the order of a . Therefore, Lagrange's theorem implies that *the order of any element of a finite group G is a divisor of the order of G* . It is also clear that if n is the order of a and $ma = 0$ for an integer m , then m must be a multiple of n . In fact, if $m = kn + r$, where $r < n$ is a remainder of the division of m by n , then $ma = (kn + r)a = k(na) + ra = 0 + ra = ra$, i.e., $ra = 0$. However, this implies that $r = 0$ (since $r < n$ and n is the smallest integer satisfying the equation $na = 0$).

Let us also note that, if we use a multiplicative group notation, then the order n of element a must be defined as the smallest integer n satisfying the relation $a^n = 1$. The cyclic subgroup generated by a consists in this case of the elements $a^1 = a, a^2 = a \times a, a^3, \dots, a^{n-1}, a^n = 1 (= a^0)$. Moreover, the equation $a^m = 1$ is valid here if and only if m is a multiple of n .

2. The other important algebraic systems are *fields* and *rings*.

A field is a set F of elements a, b, c, \dots for which two operations are defined, associating with a pair of elements a and b of F a third element. These opera-

tions are called 'addition' (the 'sum' of elements a and b of F is written $a + b$) and 'multiplication' (the product of elements a and b is naturally denoted by ab). In addition,

F1 : the elements of a field must form a group with respect to addition;

F2 : the nonzero elements of a field must form a group with respect to multiplication;

F3 : the addition and multiplication must obey the distributive law

$$(a + b)c = ac + bc \text{ for all } a, b \text{ and } c.$$

It is easy to understand that for any elements a and b of a field F , where b is different from null element 0, there exists their 'quotient' a/b , i.e., a number y such that $by = a$; this y can be defined by the formula $y = ab^{-1}$. Moreover, it is clear that if 0 is the null element of a field (i.e., an identity element of the corresponding additive group), then $a0 = 0$ for every a (since $0 = 1 - 1 = 1 + (-1)$, and, therefore, $a0 = a \times [1 + (-1)] = a1 + a(-1) = a - a = 0$). It is also important to note that if $ab = 0$, then at least one of the elements a and b is necessarily equal to 0. In fact, if (say) $b \neq 0$, then the multiplication of the equality $ab = 0$ by b^{-1} yields $abb^{-1} = 0b^{-1}$, i.e., $a1 = 0$ or $a = 0$.

Examples

A. It is obvious that a set of all rational (or real, or complex) numbers forms a field with respect to the operations of ordinary addition and multiplication.

B. The product of numbers a and b of q -arithmetic is defined as the remainder after division of the ordinary product ab by q ; thus, say, the product of numbers a and b of a 10-arithmetic is just the last digit of the number ab . The multiplication tables for numbers in a 2-arithmetic, 5-arithmetic and 6-arithmetic assume the form of Tables 5, 6 and 7.

TABLE 5

\times	0	1
0	0	0
1	0	1

TABLE 6

\times	0	1	2	3	4
0	0	0	0	0	0
1	0	1	2	3	4
2	0	2	4	1	3
3	0	3	1	4	2
4	0	4	3	2	1

TABLE 7

\times	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	1	2	3	4	5
2	0	2	4	0	2	4
3	0	3	0	3	0	3
4	0	4	2	0	4	2
5	0	5	4	3	2	1

A comparison of these tables enables us to make a salient distinction among them : whereas for 2-arithmetic and 5-arithmetic each row of the table, except the first one, which has all zeros, contains a one, it is not so for 6-arithmetic

(here, the 1st, 3rd, 4th and 5th rows do not contain a one). Thus, in 2-arithmetic and 5-arithmetic, *every number different from 0 has an inverse* (in 2-arithmetic we have $1^{-1} = 1$; in 5-arithmetic we have the equalities $1^{-1} = 1$, $2^{-1} = 3$, $3^{-1} = 2$ and $4^{-1} = 4$); on the contrary, in a 6-arithmetic, the numbers 2, 3 and 4 have no inverse. Hence, it easily follows that 2-arithmetic and 5-arithmetic defined with respect to addition and multiplication *are fields*, but 6-arithmetic *is not* a field.

It is trivial to see that for every *composite* $q = kl$ (with $k > 1$, $l > 1$) a q -arithmetic *cannot* form a field: this stems, for example, from the fact that here $kl = 0$ (where multiplication is understood in the sense indicated above). If, however, p is a *prime* number, then in a p -arithmetic every number has an inverse (see p. 376 below); hence a p -arithmetic with the operations of addition and multiplication of numbers defined in it *is a finite field* F_p of p elements (or a field of *order* p).

A description of all possible finite fields shall be given later in this appendix. It is, however, convenient to consider now some general properties of finite fields. We know that all elements of a field form an additive group and all its non-zero elements form a multiplicative group. If F_N is a finite field of order N , then the corresponding additive group is of the same order N and the multiplicative group is of order $N - 1$. Every element a of F_N has an additive order n_1 equal to the smallest integer satisfying the relation $n_1 a = 0$. If $a \neq 0$, then it has also a multiplicative order n_2 equal to the smallest integer satisfying the relation $a^{n_2} = 1$. According to the general results stated above, the integer n_1 is necessarily the divisor of N and the integer n_2 is the divisor of $N - 1$.

The additive order of the unit element 1 (i.e., the smallest integer n satisfying the relation $n1 = 0$) is called the *characteristic* of F_N . It is easy to show that the characteristic n is necessarily a prime. In fact, if k and l are two arbitrary integers, then, evidently,

$$\begin{aligned} (k1) \times (l1) &= \underbrace{(1 + 1 + \dots + 1)}_{k \text{ terms}} \times \underbrace{(1 + 1 + \dots + 1)}_{l \text{ terms}} \\ &= \underbrace{1 + 1 + \dots + 1}_{kl \text{ terms}} = kl1. \end{aligned}$$

If now $n = kl$, where $k \neq n$ and $l \neq n$, then $n1 = kl1 = (k1) \times (l1) = 0$, but $k1 \neq 0$ and $l1 \neq 0$ by virtue of the definition of the characteristic n . Since this is impossible, n must be a prime. Therefore, it is reasonable to denote hereafter the characteristic by the letter p which is commonly used to write a prime. An example of a field having characteristic p is, of course, given by p -arithmetic.

We know that the order N of a field F must be a multiple of its characteristic p . Later, it will be elucidated that N must have the form p^k , where k is an

integer. In the coding theory related to the binary communication channels the fields of characteristic 2 are most important. Every such field consists of 2^k elements.

The multiplicative order n of an element a of a field F_N must be a divisor of $N - 1$. The element a of a multiplicative group of order $N - 1$ is called a *primitive element* of F_N . If a is a primitive element of F_N , then the elements $a, a^2, a^3, \dots, a^{N-2}, a^{N-1} = 1$ coincide with all nonzero elements of F_N . In other words, a multiplicative group of F_N is a cyclic group generated by the primitive element a .

Let us now prove the following important assertion: *any finite field F_N contains a primitive element a* . Consider at first two elements b_1 and b_2 having relatively prime (multiplicative) orders n_1 and n_2 . Then, it is easy to show that the order of $b_1 b_2$ is equal to $n_1 n_2$. In fact, if $(b_1 b_2)^k = 1$, then

$$b_1^k = b_2^{-k}, b_1^{n_1 k} = b_2^{-n_1 k} = (b_2^{n_2})^{-k} = 1,$$

and, similarly, $b_2^{n_2 k} = 1$. Therefore, $n_2 k$ must be a multiple of n_1 and $n_1 k$ must be a multiple of n_2 . Since n_1 and n_2 are relatively prime numbers, k must be a multiple of $n_1 n_2$. Moreover,

$$(b_1 b_2)^{n_1 n_2} = (b_1^{n_1})^{n_2} (b_2^{n_2})^{n_1} = 1 \times 1 = 1,$$

and hence $n_1 n_2$ is equal to the order $b_1 b_2$.

Let us now consider an element a of F_N having the highest order n and suppose that $n < N - 1$. All elements of F whose orders are divisors of n satisfy the equation $x^n = 1$, i.e., $x^n - 1 = 0$. It is easy to show that any equation of order n with the coefficients from an arbitrary field F cannot have more than n distinct roots in the field F . (The proof of this statement is completely analogous to the proof of the well-known special case of it related to the field F of real numbers.) Since $n < N - 1$, all nonzero elements of F_N cannot be the roots of equation $x^n - 1 = 0$. Hence, the field F_N contains at least one element b whose order m is not a divisor of n . Let us assume that $m = kl$, where l is the greatest common divisor of m and n but k and n are relatively prime numbers. Since b has order m , it is clear that b^l has order k and, consequently, ab^l has order nk . But this contradicts the assumption of n being the highest order of all nonzero elements of F_N . Therefore, $n = N - 1$, i.e., a is a primitive element of F_N .

If a field F_N has characteristic p , then, evidently, $pa = a + a + \dots + a$ (p summands!) is equal to zero for every element a of F_N . In fact, $a = a \times 1$ and $pa = a + a + \dots + a = (a \times 1) + (a \times 1) + \dots + (a \times 1) = a(1 + 1 + \dots + 1) = a \times 0 = 0$. In particular, $2a = 0$ for every element a of a field of characteristic 2. Therefore, we have

$$(a + b)^2 = (a + b)(a + b) = a^2 + 2ab + b^2 = a^2 + b^2, \quad (1)$$

$$(a + b + c)^2 = [(a + b) + c]^2 = (a + b)^2 + c^2 = a^2 + b^2 + c^2, \quad (2)$$

and so on, so that

$$(a_1 + a_2 + \dots + a_m)^2 = a_1^2 + a_2^2 + \dots + a_m^2. \quad (3)$$

Quite similarly it can be shown that in a field of characteristic p

$$(a_1 + a_2 + \dots + a_m)^p = a_1^p + a_2^p + \dots + a_m^p. \quad (3a)$$

We shall now consider a simple statement which is used in the construction of certain error-correcting codes. Suppose that a and b are two distinct nonzero elements of any (finite or infinite) field F . It is easy to show that the equations

$$ax + by = 0, \quad a^2x + b^2y = 0, \quad (4)$$

where x and y are also elements of F , imply that $x = 0$ and $y = 0$. In fact, if we multiply the first equation by b and then subtract from it the second equation, we obtain

$$abx - a^2x = 0, \text{ i.e., } a(b - a)x = 0,$$

and hence $x = 0$ (since $a \neq 0$ and $b - a \neq 0$). Then, of course, $y = 0$, too. Moreover, if a , b and c are three distinct nonzero elements of F , then the following three equations

$$ax + by + cz = 0, \quad a^2x + b^2y + c^2z = 0, \quad a^3x + b^3y + c^3z = 0 \quad (4a)$$

imply that $x = y = z = 0$. In fact, if we multiply the first equation by c and subtract from it the second equation, and also multiply the second equation by c and subtract from it the third equation, then we obtain

$$a(c - a)x + b(c - b)y = 0, \quad a^2(c - a)x + b^2(c - b)y = 0.$$

But these equations have the same form as equations (4), only x and y are now replaced by $(c - a)x$ and $(c - b)y$. Therefore, $(c - a)x = (c - b)y = 0$ and, consequently, $x = y = 0$ and also $z = 0$.

The same arguments can be applied to the case of four similar equations, and so on. By mathematical induction we may conclude that if a_1, a_2, \dots, a_m are m distinct nonzero elements of a field F and

$$a_1x_1 + a_2x_2 + \dots + a_mx_m = 0,$$

Examples

(a) It is plain that a field is a special case of a ring (a field is a ring with division); hence *all examples of a field are simultaneously also examples of a ring*.

(b) *A set of all integers forms a ring* (with respect to the operations of ordinary addition and multiplication over numbers).

(c) *A collection of all polynomials with coefficients in some field F forms a ring* with respect to termwise addition and multiplication of polynomials : if

$$a(x) = a_0 + a_1x + a_2x^2 + \dots + a_{n-1}x^{n-1},$$

and

$$b(x) = b_0 + b_1x + b_2x^2 + \dots + b_{m-1}x^{m-1},$$

then

$$a(x)b(x) = a_0b_0 + (a_0b_1 + a_1b_0)x + (a_0b_2 + a_1b_1 + a_2b_0)x^2 + \dots + a_{n-1}b_{m-1}x^{n+m-2}.$$

The null element of this ring is the polynomial 0, and the unit element is the polynomial 1 (both the polynomials are of zero degree).

Examples (b) and (c) share, in fact, many common properties. One, for example, is the existence in both the rings considered of *division with a remainder term* of the number a by b or the polynomial $a(x)$ by $b(x)$ (where $|a| \geq |b|$ and $\deg a(x) \geq \deg b(x)$, respectively; by $\deg f(x)$ we denote the *degree* of the polynomial $f(x)$). This division is represented, respectively, by the equations $a = ub + r$, where $|r| < |b|$ and $a(x) = u(x)b(x) + r(x)$, where $\deg r(x) < \deg b(x)$. Here the number u (resp. the polynomial $u(x)$) is the *quotient* of the division of a by b (or $a(x)$ by $b(x)$), and the number r (the polynomial $r(x)$) is the *remainder* (the remainder of division may turn out to be 0).

The procedure of division with a remainder can be used to find the *greatest common divisor* (gcd) of two numbers or polynomials. Thus, for instance, restricting ourselves to the case of *positive integers* a and b and denoting by (a, b) the gcd of these numbers, we find consecutively:

$$a = ub + r, \quad \text{where } r < b \quad \text{and} \quad (a, b) = (b, r);$$

$$b = u_1r + r_1, \quad \text{where } r_1 < r \quad \text{and} \quad (b, r) = (r, r_1);$$

$$r = u_2r_1 + r_2, \quad \text{where } r_2 < r_1 \quad \text{and} \quad (r, r_1) = (r_1, r_2);$$

.....

$$r_{k-2} = u_k r_{k-1} + r_k, \quad \text{where } r_k < r_{k-1} \quad \text{and} \quad (r_{k-2}, r_{k-1}) = (r_{k-1}, r_k);$$

$$r_{k-1} = u_{k+1} r_k \quad \text{and, hence,} \quad (r_{k-1}, r_k) = r_k.$$

Thus, r_k is also the number

$$d = (a, b).$$

It is important to note that the number $d = (a, b)$, determined by the indicated method†, can be represented in terms of the original numbers a and b as

$$d = Ma + Nb, \quad (*)$$

where M and N are some *integers*. In fact, from the equations set forth above we successively obtain

$$\begin{aligned} r &= 1 \times a + (-u) \times b \quad (= m \times a + n \times b), \\ r_1 &= 1 \times b + (-u_1) \times r \quad (= m_1 \times a + n_1 \times b), \\ r_2 &= 1 \times r + (-u_2) \times r_1 \quad (= m_2 \times a + n_2 \times b), \dots, \\ r_k &= 1 \times r_{k-2} + (-u_k) \times r_{k-1} \quad (= M \times a + N \times b), \end{aligned}$$

where all the numbers m and n (i.e., 1 and $-u$), m_1 , and n_1 (they are equal to $-u_1$ and $1 + uu_1$), m_2 and n_2 , . . . , M and N are integers.

From the formula (*) it follows in particular that in p -arithmetic (where p is a *prime*) every number $a \neq 0$ has an inverse. In fact, if $0 < a < p$, then obviously $(a, p) = 1$, and hence

$$1 = (a, p) = Ma + Np.$$

Thus, the product $Ma (= (-N) \times p + 1)$ when divided by p yields the remainder 1. But this also implies that a number m of a p -arithmetic corresponding to M (the remainder from the division of M by p) is the inverse of a : in multiplying numbers by p -arithmetic rules we have $ma = 1$ and, hence, $m = a^{-1}$.

Exactly the same procedure enables us to find the gcd of $(a(x), b(x))$ of two polynomials $a(x)$ and $b(x)$ and show that if $(a(x), b(x)) = d(x)$, then

$$d(x) = M(x) \times a(x) + N(x) \times b(x), \quad (**)$$

where $M(x)$ and $N(x)$ are some polynomials.

The analogy between a ring of integers and a ring of polynomials (with coefficients from any field F) can be characterized differently also. A subset J of elements of an arbitrary ring K is called an *ideal* of this ring, if

†This procedure of determining the gcd of a and b is called *the Euclidean division algorithm*; a ring, in which this procedure is valid (in particular, a ring of integers or a ring of polynomials) is sometimes called a *Euclidean ring*.

- (i) J is a subgroup with respect to the addition operation in K ;
- (ii) for each a in J all products ak , where k is some element of K , also belong to J .

A typical example of the ideal of a ring of integers is a set of all numbers divisible by an arbitrarily chosen integer i (i.e. of all numbers of the form ai , where a runs through *all* integral values). In analogy to this, an example of an ideal in a set of polynomials is a set of polynomials that are divisible by an arbitrary preassigned polynomial $i(x)$ (i.e., a set of polynomials of the form $a(x)i(x)$, where $a(x)$ is an *arbitrary* polynomial). An ideal so constructed is called a *principal ideal* of a ring of integers (resp. polynomials) generated by a number i (resp. polynomial $i(x)$).

We make the following statement that reveals the deeper combined characteristics of rings of integers and polynomials.

In a ring of integers or in a ring of polynomials every ideal J is a principal ideal, i.e., it consists of all possible multiples of a fixed integer i (corresp. polynomial $i(x)$).

The proof of this statement presents no difficulty. Indeed, it is as a matter of fact possible that an ideal of a ring of *integers* consists of just a single number 0 (for this one-element set all the conditions defining an ideal are obviously satisfied), but in such a case this is a principal ideal generated by the number 0. If, however, this is not so, then we denote by i a *least number in absolute magnitude*, different from zero, that belongs to an ideal J (for simplicity we may agree to consider, say, that $i > 0$). It is now required to show that every other non-zero number b belonging to J is necessarily a *multiple* of i . Since $|b| \geq i$, b can be partitioned by i :

$$b = ai + r, \quad \text{where } 0 \leq r < i.$$

However, since J is an ideal, together with b and i , the numbers ai , $-ai$ and $r = b + (-ai)$ also belong to it. Hence $r = 0$ (because i is a *least* number in absolute magnitude, different from zero, belonging to J) and, hence, $b = ai$.

For the case of a *ring of polynomials* the statement is proved in exactly the same way; here it is necessary only to take $i(x)$ as a nonzero polynomial of *least degree*, belonging to the ideal J .

We now pass on to further examples of a ring.

(d) We have already seen that a q -*arithmetic* with addition and multiplication defined for it is a *ring of q elements* (a finite ring or a ring of *finite order* q). Moreover, if q is *prime*, then our ring is a field.

(e) We noted above that a collection of polynomials of degree $< n$, where n is a fixed number, is a group with respect to addition (a finite group, if the coefficients of the polynomials are elements of a finite field). However, such polynomials do not form a ring because the degree of the product of two

polynomials is in general *higher* than the degree of either of the factors. In order to transform a collection of polynomials of degree $< n$ into a ring, we may proceed as follows.

We choose a fixed (any one convenient to us) n th degree polynomial $Q(x)$ and agree to *replace every polynomial by the remainder of its division by $Q(x)$* ; the degree of this remainder is then obviously $< n$. Thus we arrive at a ' $Q(x)$ -arithmetic' of polynomials, which contains no polynomial of degree $\geq n$. In particular, the 'product' of two polynomials, understood in the sense of ' $Q(x)$ -arithmetic', always has degree $< n$. A $Q(x)$ -arithmetic is always (i.e., for any choice of the polynomial $Q(x)$) a ring: it is a finite ring, if the field of coefficients of polynomials is finite. If the field F of coefficients is of order p and $\deg Q(x) = n$, then the ring under consideration is of order p^n .

We give below multiplication Tables 8 and 9 for four polynomials of degree < 2 with coefficients from a 2-arithmetic in $(x^2 + x)$ -arithmetic and $(x^2 + x + 1)$ -arithmetic.

TABLE 8

\times	0	1	x	$x + 1$
0	0	0	0	0
1	0	1	x	$x + 1$
x	0	x	x	0
$x + 1$	0	$x + 1$	0	$x + 1$

TABLE 9

\times	0	1	x	$x + 1$
0	0	0	0	0
1	0	1	x	$x + 1$
x	0	x	$x + 1$	1
$x + 1$	0	$x + 1$	1	x

A comparison of these two tables is instructive. The last two rows of Table 8 *do not contain* the number 1, implying that in $(x^2 + x)$ -arithmetic the polynomials x and $x + 1$ have no inverse. In contrast, in Table 9 all rows contain 1, except the single first row which consists only of zeros. This means that in $(x^2 + x + 1)$ -arithmetic *all* polynomials different from zero have an inverse:

$$1^{-1} = 1, \quad x^{-1} = x + 1 \quad \text{and} \quad (x + 1)^{-1} = x.$$

Thus, whereas an $(x^2 + x)$ -arithmetic of polynomials with coefficients from a 2-arithmetic is only a *ring*, an $(x^2 + x + 1)$ -arithmetic of polynomials with coefficients from the same field forms a *field*. It is not difficult to comprehend the reason for this distinction. The polynomial $Q(x) = x^2 + x$ is *composite*, it is partitioned into two factors of degree one:

$$x^2 + x = x(x + 1).$$

This implies that $(x^2 + x)$ -arithmetic cannot form a field (this is, for example, implied by the fact that here $x(x + 1) = 0$). On the contrary, the polynomial $P(x) = x^2 + x + 1$ is *prime* (or, as we often say in algebra, is *irreducible*); it cannot be partitioned into factors of degree ≥ 1 . This, in turn, directly implies that in a $P(x)$ -arithmetic every polynomial $a(x) \neq 0$ has an inverse; the proof of this fact is based on formula (**) on p. 376 and is quite similar to the proof of the fact that in a p -arithmetic, where p is *prime*, every number a has an inverse.

Thus we arrive at one more example of a field.

C. If $P(x)$ is an *irreducible* polynomial with coefficients from a certain field F , then the $P(x)$ -arithmetic with coefficients in F forms a field. If F is a finite field F_p described above of order p (where p is an arbitrary *prime* number) and $\deg P(x) = n$, then the field obtained is of order p^n .

$P(x)$ -arithmetic is not only an important example of a field — it also admits a different interpretation. Let us recall the formation of a field of complex numbers F_c by an extension of a field of real numbers F_r . It is a basic fact that the equation $x^2 + 1 = 0$ cannot be solved within the field F_r . To make this equation solvable we extend the field F_r by adding a new element i that denotes a (non-existing) root of the considered equation. In other words, we agree that $i^2 + 1 = 0$. The field F_c that contains all real numbers and also the element i must obviously contain all binomials $a + bi$, a and b being arbitrary real numbers. However, the powers of i can be easily eliminated; we know that $i^2 + 1 = 0$ and, therefore, an arbitrary polynomial $T(i)$ in i with real coefficients can be replaced by the remainder of the division of $T(i)$ by $i^2 + 1$. In fact, if $T(i) = t(i)(i^2 + 1) + r(i)$, then $T(i) = r(i)$ within the field F_c . But $r(i)$ is also a binomial of the form $a + bi$. Therefore, our field F_c consists of all binomials $a + bi$ with the usual addition and multiplication supplemented by the rule that every polynomial in i must be replaced by its remainder after division by $i^2 + 1$ (i.e., i^2 must be replaced by -1). This construction leads to the customary *field of complex numbers* which is, of course, equivalent to $(x^2 + 1)$ -arithmetic.

A quite similar procedure can be applied to obtain a new interpretation of the arbitrary $P(x)$ -arithmetic. Suppose that $P(x)$ is an irreducible n th degree polynomial with coefficients in a field F . Then, the equation $P(x) = 0$ is, of

course, insolvable within F (otherwise $P(x)$ would be reducible). Let us extend F such that this equation becomes solvable within the extended field F^* . For this, we must add to the field F a symbol α which is the root of the considered equation (so that by definition $P(\alpha) = 0$). Since F^* is a field, it must include also all polynomials in α with the coefficients in F . However, in the case in which the degree of the polynomial $T(\alpha)$ exceeds n this polynomial can be replaced within F^* by the remainder after division of $T(\alpha)$ by $P(\alpha)$. Therefore, we can consider that the field F^* consists of all polynomials of the form $a_0 + a_1\alpha + \dots + a_{n-1}\alpha^{n-1}$, where a_0, a_1, \dots, a_{n-1} are the elements of F , with the usual addition and multiplication supplemented by the replacement of the obtained product by the remainder of its division by $P(\alpha)$. It is clear that the field F^* is equivalent to $P(x)$ -arithmetic.

It can be shown that for every prime p and for each $k > 1$ there exist irreducible k th degree polynomials with coefficients from a field F_p ; hence, it follows that *for every integer $k \geq 1$ and every prime p there exists a finite field of order p^k* (a field of order $p^1 = p$ is formed by a p -arithmetic itself). Moreover, although there may exist many irreducible polynomials $P(x)$ of a given degree k with coefficients from a field F_p , all $P(x)$ -arithmetics corresponding to them are constructed alike: for every prime p and each $k \geq 1$ there exists just a *single* (to within the rearrangement of elements) field of order p^k . If, however, the integer m does not assume the form p^k (i.e., if m contains at least two *distinct* prime factors), then a field of order m *does not exist* at all.†

Before we conclude we note further that since the $Q(x)$ -arithmetic is obtained from a ring of all polynomials (with coefficients in some chosen field F) by 'coalescing' all polynomials that yield one and the same remainder when divided by $Q(x)$, the *ideals* of a $Q(x)$ -arithmetic are also obtained from the ideals of a ring of all polynomials by similarly identifying all those polynomials of an ideal that yield the same remainder on division by $Q(x)$. This, in turn, implies that the ideals of a $Q(x)$ -arithmetic are constructed in the same manner as are the ideals of a ring of all polynomials: here also *every ideal is a principal ideal* (i.e., it consists of all polynomials that are multiples in the sense of the $Q(x)$ -arithmetic of some fixed polynomial $i(x)$). However, in this connection it is necessary to keep in mind, as is easy to perceive from formula (**) on p. 376, that a set of all those polynomials taken in the sense of a $Q(x)$ -arithmetic that are multiples of a given polynomial $i(x)$ coincides with a set of all polynomials that are multiples of a polynomial $d(x)$, where $d(x) = (Q(x), i(x))$ is the gcd of the polynomials $i(x)$ and $Q(x)$. Hence, it follows that, for an *irreducible* (prime) polynomial $Q(x)$, a $Q(x)$ -arithmetic contains no ideal other than 0 and

†Thus, a field of finite order m exists if $m = p^k$, where p is some prime number, but does not exist for all other m numbers. Moreover there is only a *single* field of order p^k for every prime p and positive integer k . All these fields are due to Évariste Galois (1811-1832), the noted French mathematician, and are hence called *Galois fields*

the entire ring (the entire $Q(x)$ -arithmetic itself); in fact, here the gcd of $Q(x)$ and $i(x)$ is either 1 or $Q(x)$. If, however, a polynomial $Q(x)$ is *reducible*, i.e., if it is divisible into factors whose degree is less than $\deg Q(x)$, then a set of all polynomials that are multiples of each factor of $Q(x)$ forms an ideal of a $Q(x)$ -arithmetic. Thus, by way of example, in the case of a $(x^2 + x)$ -arithmetic over a 2-arithmetic a set of all ideals consists of 'zero ideals' $\{0\}$; the entire $(x^2 + x)$ -arithmetic; a set $\{x, 0\}$ of polynomials that are multiples of x , and a set $\{x + 1, 0\}$ of polynomials that are multiples of $x + 1$ (see Table 8 on p. 378).

3. We shall now enunciate one more algebraic concept that is found to be useful in coding theory.

A set V of elements a, b, c, \dots (called vectors) forms a vector space over a field F (the elements of a field are called numbers; the null and unit elements of a field are denoted below by the symbols 0 and 1), if

- (i) *for the set of vectors the operation of addition is defined such that the vectors form a group (the null element of this group is denoted by 0);*
- (ii) *the operation of multiplication of a vector by a number is defined; moreover, the product aa (where a is a number and a a vector) is a vector; and*

VS1 : *the multiplication of a vector by a number is associative : $a(ba) = (ab)a$ for all numbers a, b and every vector a ;*

VS2 : *the multiplication of a vector by a number is distributive relative to the addition of numbers :*

$$(a + b)a = aa + ba \text{ for all numbers } a, b \text{ and every vector } a;$$

VS3 : *the multiplication of a vector by a number is distributive relative to the addition of vectors :*

$$a(a + b) = aa + ab \text{ for every number } a \text{ and all vectors } a, b;$$

VS4 : $1a = a$ *for every vector a .*

From the properties (axioms) of multiplication of a vector by a number it is easy to deduce also that

$0a = 0$ for every vector a ; $a0 = 0$ for every number a ; $(-1)a = -a$ for every vector a .

Examples

A. *Blocks (or vectors) $a = (a_0, a_1, \dots, a_{N-1})$, where N is a fixed natural number and a_0, a_1, \dots, a_{N-1} are arbitrary numbers in a field F , form a vector space with respect to the operations of addition of vectors and the multiplication*

of a vector by a number, which are defined as follows : if

$$\mathbf{a} = (a_0, a_1, \dots, a_{N-1}) \quad \text{and} \quad \mathbf{b} = (b_0, b_1, \dots, b_{N-1}),$$

then

$$\mathbf{a} + \mathbf{b} = (a_0 + b_0, a_1 + b_1, \dots, a_{N-1} + b_{N-1});$$

if $\mathbf{a} = (a_0, a_1, \dots, a_{N-1})$, then $a\mathbf{a} = (aa_0, aa_1, \dots, aa_{N-1})$. Here F is called a *field of scalars* or a *basic field*, over which the vector space V is constructed; the numbers a_0, a_1, \dots, a_{N-1} are called the *coordinates* of the vector \mathbf{a} and the number N the *dimension* of V .

If F is an infinite field, then the number of all possible vectors is also infinite; if F is, however, of order m , then the vector space V of dimension N (N -dimensional vector space) contains only m^N distinct vectors.

This example is central; the other examples always reduce to it.

B. The vectors (directed segments) of a plane or the usual (three-dimensional) space form a vector space with respect to the operations of addition of vectors and the multiplication of a vector by a (real) number, defined as follows :

$$\overline{OA} + \overline{OB} = \overline{OC}$$

if OC is a diagonal of the parallelogram $OACB$, constructed on the segments OA and OB ;

$$\overline{OD} = a \times \overline{OA}$$

if the segments \overline{OD} and \overline{OA} belong to the same straight line;

$$OD = |a| \times OA$$

and D and A lie on the same side with respect to O if $a > 0$, but on opposite sides if $a < 0$.

Example **B** reduces to the central Example **A** if we introduce in the usual manner the coordinates (x, y) of the vector \overline{OA} of a plane (Fig. 38a) and the coordinates (x, y, z) of the vector \overline{OA} of a space (Fig. 38b). It is also found that in the case of vectors of a plane, if $\mathbf{a} = (x, y)$ and $\mathbf{b} = (x_1, y_1)$, then

$$\mathbf{a} + \mathbf{b} = (x + x_1, y + y_1) \quad \text{and} \quad a\mathbf{a} = (ax, ay);$$

in the case of vectors of a space, if $\mathbf{a} = (x, y, z)$ and $\mathbf{b} = (x_1, y_1, z_1)$, then

$$\mathbf{a} + \mathbf{b} = (x + x_1, y + y_1, z + z_1) \quad \text{and} \quad a\mathbf{a} = (ax, ay, az).$$

Thus, the vectors of a plane form a two-dimensional vector space, and the vectors of a space—a three-dimensional vector space over a field of real numbers.

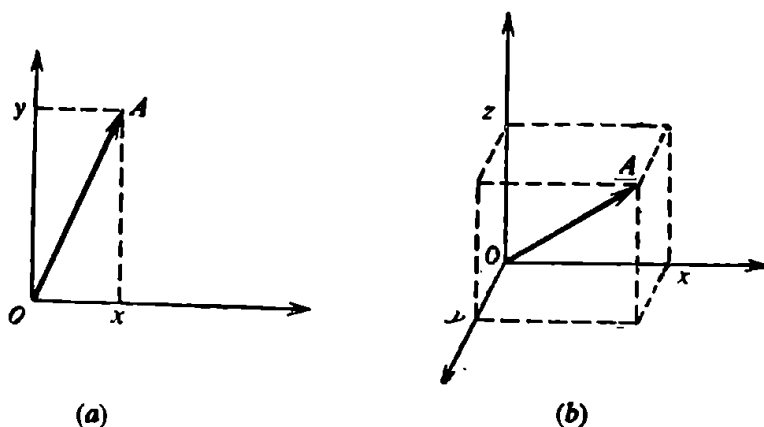


Fig. 38.

C. It is obvious that a set of all $(m \times n)$ -matrices with elements in field F forms an (mn) -dimensional vector space over F if the addition of matrices is defined as above, and the multiplication of a matrix by a number a is defined as the multiplication of all elements of the matrix by this number. This example differs from the basic Example A only in that here the mn coordinates of a vector are written not in a single row, but in the form of a rectangular matrix.

D. All polynomials of degree less than n

$$a_0 + a_1x + a_2x^2 + \dots + a_{n-1}x^{n-1}$$

with coefficients in a field F form an n -dimensional vector space over F . In fact, every polynomial can be characterized by its coefficients a_0, a_1, \dots, a_{n-1} (which, if considered more convenient, can be set out within parentheses), and the (ordinary) addition of polynomials and the multiplication of a polynomial by a number reduce, respectively, to addition of the coefficients of two polynomials and multiplication of the coefficients of the polynomial by a number.

Since $P(x)$ -arithmetic, where $P(x)$ has degree n , consists of all polynomials of degree less than n , it is clear that $P(x)$ -arithmetic also forms an n -dimensional vector space over the coefficient field F . Note that the field F is, in fact, a collection of all polynomials of degree zero (i.e., constants) and hence F is a subfield of $P(x)$ -arithmetic. It is also possible to prove that, if F is an arbitrary field and F_0 is its subfield (i.e., a set of the elements of F forming a field with respect to the operations defined in F), then F necessarily forms a vector space over the field F_0 . A finite field having characteristic p includes evidently a subfield F_0 consisting of p elements $1, 1 + 1 = 2, 2 + 1 = 3, \dots, (p - 1) + 1 = 0$. Therefore, the field F can be represented as a vector space over F_0 and, of course, this vector space must be finite-dimensional. This fact implies the result stated above that a finite field of characteristic p must have p^k distinct elements.

E. All possible polynomials

$$a_0 + a_1x + a_2x^2 + \dots + a_kx^k$$

(their degrees are now unrestricted) also form a vector space with respect to the operations of ordinary addition of polynomials and the multiplication of a polynomial by a number. This example, however, does not coincide with Example A, because the number of coefficients of a polynomial can be arbitrarily large; hence we say that a space of all polynomials does not have a dimension (sometimes we say differently that it has an *infinite* dimension).

We now suppose that W is some portion of the vectors of a vector space V . If this set W satisfies the following conditions :

SS1 : if the vectors a, b belong to W , then $a + b$ also belongs to W ;

SS2 : if a belongs to W , then all vectors aa also belong to W , where a are all possible numbers,

then W is itself a vector space with respect to the operations (defined in V) of addition of the vectors and the multiplication of a vector by a number. In this case we say that W is a (*linear* or *vector*) *subspace* of a vector space V .

In particular, if V is a set of vectors \overline{OA} of an ordinary space, and W is a plane that passes through the point O (Fig. 39), then the vectors \overline{OB} belonging

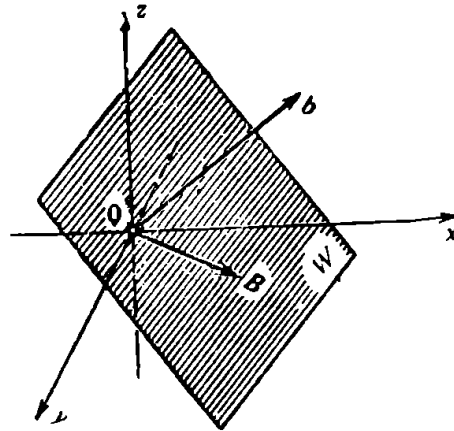


Fig. 39.

to W form the subspace of a three-dimensional vector space. If V is a set of all n -dimensional vectors

$$a = (a_1, a_2, \dots, a_n),$$

then a subspace is formed by a set W of vectors a , whose coordinates satisfy a fixed equation of the form

$$b_1 a_1 + b_2 a_2 + \dots + b_n a_n = 0, \quad (5)$$

where b_1, b_2, \dots, b_n are arbitrary fixed 'numbers', i.e., are elements of the field,

p is prime; however, in the case of a base field different from p -arithmetic (say, when $P(x)$ -arithmetic figures as a base field, where $P(x)$ is an irreducible polynomial), there exist subgroups of a vector space such that they are not its subspaces.

The notion of linear dependence of vectors is related to the notion of linear subspace of a vector space. The vectors a_0, a_1, \dots, a_n of a vector space V are said to be *linearly dependent*, if there are numbers (i.e., the elements of the basic field F) $\lambda_0, \lambda_1, \dots, \lambda_n$ such that not all of them are equal to zero and

$$\lambda_0 a_0 + \lambda_1 a_1 + \dots + \lambda_n a_n = 0. \quad (7)$$

Conversely, if the numbers $\lambda_0, \lambda_1, \dots, \lambda_n$ satisfying the above conditions do not exist, the vectors a_0, a_1, \dots, a_n are said to be *linearly independent*. It is easy to show that, if a_1, \dots, a_n are n linearly independent vectors of a vector space V , then a set of all the vectors a_0 such that $n + 1$ vectors a_0, a_1, \dots, a_n are linearly dependent forms an n -dimensional linear subspace of V . In fact, it is clear that the coefficient λ_0 of equation (7) is here necessarily different from zero for all vectors a_0 (since, otherwise, the vectors a_1, \dots, a_n would be linearly dependent). Now, if we multiply equation (7) by λ_0^{-1} and write

$$\mu_i = -\lambda_i \lambda_0^{-1} = \frac{\lambda_i}{\lambda_0},$$

we obtain

$$a_0 = \mu_1 a_1 + \dots + \mu_n a_n. \quad (8)$$

Therefore, a set of vectors a_0 , satisfying the condition that a_0, a_1, \dots, a_n are linearly dependent vectors coincides with a set of vectors of the form (8), where the coefficients μ_1, \dots, μ_n run through the field F . It is clear that the last vector set satisfies conditions SS1 and SS2 which define a linear subspace. Moreover, a 'block' of n numbers $m = (\mu_1, \mu_2, \dots, \mu_n)$ can be associated with every vector a_0 , while two different vectors $a_0^{(1)}$ and $a_0^{(2)}$ correspond to any pair of distinct blocks $m^{(1)}$ and $m^{(2)}$. (In fact, if

$$a_0 = \mu_1^{(1)} a_1 + \dots + \mu_n^{(1)} a_n = \mu_1^{(2)} a_1 + \dots + \mu_n^{(2)} a_n,$$

where not all the differences $\mu_i^{(1)} - \mu_i^{(2)}$ are equal to zero, then $\mu_1 a_1 + \dots + \mu_n a_n = 0$, i.e., the vectors a_1, \dots, a_n are linearly dependent, contradicting our assumption.) This shows that a linear subspace of all vectors a_0 is really n -dimensional. (In particular, if F is a finite field of order q , then the number of all blocks m and, therefore, also the number of vectors a_0 is equal to q^n . This again shows that the considered subspace is of dimension n .)

It follows from the result stated that *every system of $(N + 1)$ vectors a_1, a_2, \dots, a_{N+1} of an N -dimensional vector space V is necessarily linearly dependent.* In fact, if it were not so, then a set of all vectors a_0 such that $a_0, a_1, a_2, \dots, a_{N+1}$ are linearly dependent would be an $(N + 1)$ -dimensional linear subspace. However, it is clear that an N -dimensional vector space V cannot have an $(N + 1)$ -dimensional linear subspace. (In particular, if the basic field F is of finite order, then the total number of all vectors of V is insufficient to form an $(N + 1)$ -dimensional subspace.) As a specific example, let us mention $P(x)$ -arithmetic over a finite field F , where $P(x)$ is an irreducible polynomial of order n . We know that this $P(x)$ -arithmetic forms an n -dimensional vector space over F . It is, therefore, clear that every system of $(n + 1)$ elements of a $P(x)$ -arithmetic must be linearly dependent.

From the concept of a vector space it is easy to pass on to the main geometric concept of *Euclidean space*. To be precise, an N -dimensional vector space E is called *Euclidean*, if in it is defined the length $|a|_E$ (or, simply $|a|$) of a vector a with coordinates $(a_0, a_1, \dots, a_{N-1})$:

$$|a|_E = \sqrt{a_0^2 + a_1^2 + \dots + a_{N-1}^2}. \quad (+)$$

(Obviously, the basic field here must be such that there exists in it a square root of the sum of the squares of any pair of elements of a field.) Further, if we agree to call the vectors of Euclidean space 'points', associating with the null vector 0 some point O and with the vector a a point A with the same coordinates, and also to write $a = \overline{OA}$, then the distance $|AB|_E$, or simply $|AB|$ between the points A and B is defined by

$$|AB| = |\overline{OB} - \overline{OA}| = \sqrt{(b_0 - a_0)^2 + (b_1 - a_1)^2 + \dots + (b_{N-1} - a_{N-1})^2}, \quad (++)$$

where $(a_0, a_1, \dots, a_{N-1})$ and $(b_0, b_1, \dots, b_{N-1})$ are coordinates of A and B (i.e., of vectors \overline{OA} and \overline{OB}). This permits us to characterize the subject-matter of Euclidean geometry as a description of those properties of figures (i.e. sets of points) in the Euclidean space E that are identical for every pair of equal figures (where the equality of two figures is defined by the condition of equality of distances between any pair of points of these two figures, corresponding to each other).

A Euclidean space with the *real* coordinates of points (and vectors) is an example of a *metric* vector space. A set M of points is called a *metric space* if for every pair of points A and B there is defined a (real) number ρ_{AB} , called the *distance* between A and B , and

- MS1 : $\rho_{AB} > 0$ for $A \neq B$, $\rho_{AA} = 0$ (*positiveness* of distance);
- MS2 : $\rho_{AB} = \rho_{BA}$ (*symmetry* of distance);
- MS3 : $\rho_{AB} + \rho_{BC} \geq \rho_{AC}$ for every A, B and C (*triangle inequality*).

If the number $\rho_{AB} = |AB|_E$ is defined by the formula $(++)$, then the conditions MS1 and MS2 are obviously satisfied. It may not be that simple to establish MS3, i.e., the inequality

$$\begin{aligned} & \sqrt{(b_0 - a_0)^2 + (b_1 - a_1)^2 + \dots + (b_{N-1} - a_{N-1})^2} \\ & + \sqrt{(c_0 - b_0)^2 + (c_1 - b_1)^2 + \dots + (c_{N-1} - b_{N-1})^2} \\ & \geq \sqrt{(c_0 - a_0)^2 + (c_1 - a_1)^2 + \dots + (c_{N-1} - a_{N-1})^2}, \end{aligned}$$

yet this does not present any singular difficulty.†

There are also many other methods of introducing a 'metric' in an N -dimensional vector space. Thus, for instance, in many respects the so-called 'Minkowski metric'†† is much simpler than the Euclidean metric (+)–(++). This 'Minkowski metric' is given by

$$|a|_M = |a_0| + |a_1| + \dots + |a_{N-1}|, \quad (\text{A})$$

and

$$|AB|_M = |b_0 - a_0| + |b_1 - a_1| + \dots + |b_{N-1} - a_{N-1}|, \quad (\text{B})$$

where $|a|$ is the absolute value of a (real) number a . Equation (B) implies directly that the distance $\rho_{AB} = |AB|_M$ also satisfies the conditions MS1–MS3.

The metric (A)–(B) can be defined for a vector space constructed over any basic field F for which there exists an *absolute value* of an element a in the field, a real number $|a|$ such that†††

- (i) $|a| > 0$ for $a \neq 0$; $|0| = 0$;
- (ii) $|ab| = |a| \times |b|$;
- (iii) $|a + b| \leq |a| + |b|$.

In particular, if the base field is a 2-arithmetic and the absolute value of an element in the field is defined by the usual equalities

$$|0| = 0, \quad |1| = 1$$

(where 0 and 1 on the right-hand sides again occur as *real numbers*), then the metric defined by equations (A) and (B) above is called the 'Hamming metric':

$$|a|_H = |a_0| + |a_1| + \dots + |a_{N-1}|,$$

$$|AB|_H = |b_0 - a_0| + |b_1 - a_1| + \dots + |b_{N-1} - a_{N-1}|.$$

It is plain that if the points $A = (a_0, a_1, \dots, a_{N-1})$ and $B = (b_0, b_1, \dots, b_{N-1})$ of an N -dimensional space with coordinates in 2-arithmetic correspond to two sequences of signals, then the distance $|AB|_H$ is equal to the number of noncoincident signals in the sequences A

†See, for example, Kuiper, N. H. (1963), *Linear Algebra and Geometry* (pp. 131–132), North Holland, Amsterdam; or Halmos, P. R. (1958), *Finite-Dimensional Vector Spaces*, § 64, Van Nostrand, Princeton.

††H. Minkowski, the German mathematician, in his researches on number theory, has considered more general methods of introducing a metric in an N -dimensional vector space, encompassing both the formulae (+) and (B).

†††The symbols 0 appearing here on the left-hand and right-hand sides of the equality $|0| = 0$ have somewhat different senses: the one on the left-hand side is an *element* of the field under consideration, the other on the right-hand side is simply a *real number*. A similar remark can be made in connection with certain other equalities below.

and B . This fact explains the usefulness of the Hamming metric in coding theory†. In addition, from the triangle inequality it follows that a pair of 'Hamming spheres' of radius n with centres Q_1 and Q_2 (i.e., a set of points A such that $|Q_1 A|_H < n$ and $|Q_2 A|_H < n$, respectively; see p. 338) cannot intersect if $|Q_1 Q_2|_H > 2n$ (we availed ourselves of this fact on p. 338).

We further note that if sequences $A(a_0, a_1, \dots, a_{N-1})$, where all a_i take the values 0 and 1, are represented by the points of an ordinary (real) N -dimensional space (these points are the vertices of a 'unit cube' of the N -dimensional Euclidean space), then obviously

$$|AB|_E = \sqrt{|AB|_H}.$$

Hence the Euclidean distance $|AB|_E$ between the points A and B , defined by formula (++), can serve as a completely satisfactory characteristic of the difference between the sequences $A(a_0, a_1, \dots, a_{N-1})$ and $B(b_0, b_1, \dots, b_{N-1})$ of the elementary signals. This position enables us to use in communication theory the results related to (N -dimensional) Euclidean geometry. In the first place, the conclusions from the so-called *discrete geometry* are here useful, since discrete geometry deals especially with the problems of 'closest packing of disjoint equal spheres' in a many-dimensional space and the problems of determining those configurations of a finite number of points located in a given domain of a space, for which the *least pairwise distance between these points is the greatest*.

In particular, the problem of determining all binary codes, where the coded messages are sequences of N elementary signals, *correcting any number of errors not exceeding n* , reduces to the problem of determining all possible fillings of a 'unit cube' of an N -dimensional Euclidean space with disjoint spheres of radius \sqrt{n} and centres at the vertices of the cube. By what has been stated, the problem of finding such fillings of an N -dimensional cube with spheres of a given radius, where the number of spheres involved is the *largest possible* (or, is at least sufficiently large), is of great interest in coding theory. However, any perspective geometric approach to the solution of this problem is unfortunately still an open problem.

4. In linear algebra, an important role is played by the operation of *matrix multiplication*, a special case of which is the *multiplication of an $(m \times n)$ -matrix by an $(n \times 1)$ -matrix (by the column vector)*:

†In the case in which a basic field F contains more than two elements, Hamming metric is defined by the same equations (A) and (B) as above, with the difference however that in the present case it is necessary to set

$$|a| = \begin{cases} 0, & \text{if } a = 0 \\ 1, & \text{if } a \neq 0. \end{cases}$$

Here the Hamming distance $|AB|_H$ is as before equal to the number of noncoincident signals in the sequences A and B .

We also remark that in coding theory, besides the 'Hamming distance,' some other metrics in a space of sequences of signals are also used. As an example, we may mention the so-called 'Lee metric', which coincides with the 'Hamming metric' in the case of a field F of two elements; however, in other cases it takes note not only of the fact that some coordinates of the points A and B do not coincide, but also of how greatly these coordinates differ from each other (see, e.g., [190], Chap. 8.2).

$$\begin{aligned}
 Ba &= \begin{bmatrix} b_{11}b_{12} \dots b_{1n} \\ b_{21}b_{22} \dots b_{2n} \\ \dots \dots \dots \\ b_{m1}b_{m2} \dots b_{mn} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \cdot \\ \cdot \\ a_n \end{bmatrix} \\
 &= \begin{bmatrix} b_{11}a_1 + b_{12}a_2 + \dots + b_{1n}a_n \\ b_{21}a_1 + b_{22}a_2 + \dots + b_{2n}a_n \\ \dots \dots \dots \\ b_{m1}a_1 + b_{m2}a_2 + \dots + b_{mn}a_n \end{bmatrix}.
 \end{aligned}$$

Obviously, in this product we can also write a vector a with coordinates a_1, a_2, \dots, a_n in the form of a row vector: $a = (a_1, a_2, \dots, a_n)$ although this is not conventional in linear algebra. It is then possible to let relations (6) on p. 385 assume the form

$$Ba = 0,$$

where 0 is the null column vector of n zeros.

For certain branches of linear algebra, we also find essential the concept of *elementary transformation* on a matrix, by which we understand here the following transformations:

- (i) the interchange of any two rows of a matrix;
- (ii) the interchange of any two columns of a matrix;
- (iii) the replacement of any row of a matrix by its sum with any other row (where the sum is understood as the row vector sum).

The matrices obtained from each other by a finite sequence of elementary transformations are called *equivalent*.

The indicated elementary transformations† are intrinsic especially to the parity-check matrix of a code. In fact, the interchange of matrix columns and rows here reduces to the renumbering of signals and checks used, respectively. However, the replacement of some row by its sum with another row implies that in place of two parity checks we check the parity of the first of the two used expressions and the sum of this expression with the second one. It is obvious that two such checks are completely equivalent to the original checks. It is also easy to establish further that by a sequence of elementary transformations each check matrix can be reduced to the form (2) on p. 318 (or equivalently, to the form differing from (2) only in that it is augmented by some

†In different problems of linear algebra, different elementary transformations are, in fact, found suitable.

additional rows made up of zeros; these rows obviously do not correspond to any new check and hence can be ignored). In fact, a zero row is of no interest; if such a row happens to exist already in the matrix, we can make it the top-most by transformation (i) and act analogously even in a case in which, in the process of transformations on a matrix, a new 'zero' row makes its appearance. We now consider the lowest row. It is clear that the element 1 appearing in it can be transferred by means of operation (ii) to the extreme right of the column. Thus adding this row to all rows in the last column of which 1 occurs and noting that in 2-arithmetic $1 + 1 = 0$, we can convert into zero all elements in the last columns, except for only a single 1 occurring in the last row. If after this the second row from below is found to consist of only zeros, we shift it upward. However, if it also contains at least a 1, then by operation (ii) we transfer it to the next to last column, and then by operation (iii) convert into zero all other elements in the next to last column. We next pass on to the third from last row and by iterating the same operations we endow the third from last column with the desired form, and so on. As a result, we obtain a matrix of form (2), possibly with only the same rows supplemented from above which include only zeros.

The applications of this result to the parity-check matrix of the code have demonstrated that *every parity-check code can be written in the form of a systematic code*, the number of parity checks in which may, however, be less than those in the original 'non-systematic' code (see p. 319 and Example on p. 336).

Appendix III

TABLE OF VALUES OF $-p \log p$

p	0	1	2	3	4	5	6	7	8	9
0.00	—	0.0100	0.0179	0.0251	0.0319	0.0382	0.0443	0.0501	0.0557	0.0612
0.01	0.0664	0.0716	0.0766	0.0815	0.0862	0.0909	0.0955	0.0999	0.1043	0.1081
0.02	0.1129	0.1170	0.1211	0.1252	0.1291	0.1330	0.1369	0.1407	0.1444	0.1486
0.03	0.1518	0.1554	0.1589	0.1624	0.1659	0.1693	0.1727	0.1760	0.1793	0.1825
0.04	0.1858	0.1889	0.1921	0.1952	0.1983	0.2013	0.2043	0.2073	0.2103	0.2132
0.05	0.2161	0.2190	0.2218	0.2246	0.2274	0.2301	0.2329	0.2356	0.2383	0.2409
0.06	0.2435	0.2461	0.2487	0.2513	0.2538	0.2563	0.2588	0.2613	0.2637	0.2661
0.07	0.2686	0.2709	0.2733	0.2756	0.2780	0.2803	0.2826	0.2848	0.2871	0.2893
0.08	0.2915	0.2937	0.2959	0.2980	0.3002	0.3023	0.3044	0.3065	0.3086	0.3106
0.09	0.3127	0.3147	0.3167	0.3187	0.3207	0.3226	0.3246	0.3265	0.3284	0.3303
0.10	0.3322	0.3341	0.3359	0.3378	0.3398	0.3414	0.3432	0.3450	0.3468	0.3485
0.11	0.3503	0.3520	0.3537	0.3555	0.3571	0.3588	0.3605	0.3622	0.3638	0.3654
0.12	0.3671	0.3687	0.3703	0.3719	0.3734	0.3750	0.3766	0.3781	0.3796	0.3811
0.13	0.3826	0.3841	0.3856	0.3871	0.3886	0.3900	0.3915	0.3929	0.3943	0.3957
0.14	0.3971	0.3985	0.3999	0.4012	0.4026	0.4040	0.4053	0.4066	0.4079	0.4092
0.15	0.4105	0.4118	0.4131	0.4144	0.4156	0.4169	0.4181	0.4194	0.4206	0.4218
0.16	0.4230	0.4242	0.4254	0.4266	0.4277	0.4289	0.4301	0.4312	0.4323	0.4335
0.17	0.4346	0.4357	0.4368	0.4379	0.4390	0.4400	0.4411	0.4422	0.4432	0.4443
0.18	0.4453	0.4463	0.4474	0.4484	0.4494	0.4504	0.4514	0.4523	0.4533	0.4543
0.19	0.4552	0.4562	0.4571	0.4581	0.4590	0.4599	0.4608	0.4617	0.4626	0.4635
0.20	0.4644	0.4653	0.4661	0.4670	0.4678	0.4687	0.4695	0.4704	0.4712	0.4720
0.21	0.4728	0.4736	0.4744	0.4752	0.4760	0.4768	0.4776	0.4783	0.4791	0.4798
0.22	0.4806	0.4813	0.4820	0.4828	0.4835	0.4842	0.4849	0.4856	0.4863	0.4870
0.23	0.4877	0.4883	0.4890	0.4897	0.4903	0.4910	0.4916	0.4923	0.4929	0.4935
0.24	0.4941	0.4947	0.4954	0.4960	0.4966	0.4971	0.4977	0.4983	0.4989	0.4994
0.25	0.5000	0.5006	0.5011	0.5016	0.5022	0.5027	0.5032	0.5038	0.5043	0.5048
0.26	0.5053	0.5058	0.5063	0.5068	0.5072	0.5077	0.5082	0.5087	0.5091	0.5096
0.27	0.5100	0.5105	0.5109	0.5113	0.5118	0.5122	0.5126	0.5130	0.5134	0.5138
0.28	0.5142	0.5146	0.5150	0.5154	0.5158	0.5161	0.5165	0.5169	0.5172	0.5176
0.29	0.5179	0.5182	0.5186	0.5189	0.5192	0.5196	0.5199	0.5202	0.5205	0.5208
0.30	0.5211	0.5214	0.5217	0.5220	0.5222	0.5225	0.5228	0.5230	0.5233	0.5235

0.31	0.5238	0.5240	0.5243	0.5245	0.5247	0.5250	0.5252	0.5254	0.5256	0.5258
0.32	0.5260	0.5262	0.5264	0.5266	0.5268	0.5270	0.5272	0.5273	0.5275	0.5277
0.33	0.5278	0.5280	0.5281	0.5283	0.5284	0.5286	0.5287	0.5288	0.5289	0.5290
0.34	0.5292	0.5293	0.5294	0.5295	0.5296	0.5297	0.5298	0.5299	0.5299	0.5300
0.35	0.5301	0.5302	0.5302	0.5303	0.5304	0.5304	0.5305	0.5305	0.5305	0.5306
0.36	0.5306	0.5306	0.5307	0.5307	0.5307	0.5307	0.5307	0.5307	0.5307	0.5307
0.37	0.5307	0.5307	0.5307	0.5307	0.5307	0.5307	0.5306	0.5306	0.5305	0.5305
0.38	0.5304	0.5304	0.5303	0.5303	0.5302	0.5302	0.5301	0.5300	0.5300	0.5299
0.39	0.5298	0.5297	0.5296	0.5295	0.5294	0.5293	0.5292	0.5291	0.5290	0.5289
0.40	0.5288	0.5286	0.5285	0.5284	0.5283	0.5281	0.5280	0.5278	0.5277	0.5275
0.41	0.5274	0.5272	0.5271	0.5269	0.5267	0.5266	0.5264	0.5262	0.5260	0.5258
0.42	0.5256	0.5255	0.5253	0.5251	0.5249	0.5246	0.5244	0.5242	0.5240	0.5238
0.43	0.5236	0.5233	0.5231	0.5229	0.5226	0.5224	0.5222	0.5219	0.5217	0.5214
0.44	0.5211	0.5209	0.5206	0.5204	0.5201	0.5198	0.5195	0.5193	0.5190	0.5187
0.45	0.5184	0.5181	0.5178	0.5175	0.5172	0.5169	0.5166	0.5163	0.5160	0.5157
0.46	0.5153	0.5150	0.5147	0.5144	0.5140	0.5137	0.5133	0.5130	0.5127	0.5123
0.47	0.5120	0.5116	0.5112	0.5109	0.5105	0.5102	0.5098	0.5094	0.5090	0.5087
0.48	0.5083	0.5079	0.5075	0.5071	0.5067	0.5063	0.5059	0.5055	0.5051	0.5047
0.49	0.5043	0.5039	0.5034	0.5030	0.5026	0.5022	0.5017	0.5013	0.5009	0.5004
0.50	0.5000	0.4996	0.4991	0.4987	0.4982	0.4978	0.4973	0.4968	0.4964	0.4959
0.51	0.4954	0.4950	0.4945	0.4940	0.4935	0.4930	0.4926	0.4921	0.4916	0.4911
0.52	0.4906	0.4901	0.4896	0.4891	0.4886	0.4880	0.4875	0.4870	0.4865	0.4860
0.53	0.4854	0.4849	0.4844	0.4839	0.4833	0.4828	0.4822	0.4817	0.4811	0.4806
0.54	0.4800	0.4795	0.4789	0.4784	0.4778	0.4772	0.4767	0.4761	0.4755	0.4750
0.55	0.4744	0.4738	0.4732	0.4726	0.4720	0.4714	0.4708	0.4702	0.4697	0.4691
0.56	0.4684	0.4678	0.4672	0.4666	0.4660	0.4654	0.4648	0.4641	0.4635	0.4629
0.57	0.4623	0.4616	0.4610	0.4603	0.4597	0.4591	0.4584	0.4578	0.4571	0.4565
0.58	0.4558	0.4551	0.4545	0.4538	0.4532	0.4525	0.4518	0.4512	0.4505	0.4498
0.59	0.4491	0.4484	0.4477	0.4471	0.4464	0.4457	0.4450	0.4443	0.4436	0.4429
0.60	0.4422	0.4415	0.4408	0.4401	0.4393	0.4386	0.4379	0.4372	0.4365	0.4357
0.61	0.4350	0.4343	0.4335	0.4328	0.4321	0.4313	0.4306	0.4298	0.4291	0.4283
0.62	0.4276	0.4268	0.4261	0.4253	0.4246	0.4238	0.4230	0.4223	0.4215	0.4207
0.63	0.4199	0.4192	0.4184	0.4176	0.4168	0.4160	0.4153	0.4145	0.4137	0.4129
0.64	0.4121	0.4113	0.4105	0.4097	0.4089	0.4080	0.4072	0.4064	0.4056	0.4048

Appendix III Table (contd. on page 394)

Appendix III Table (contd. from page 393)

P	0	1	2	3	4	5	6	7	8	9
0.65	0.4040	0.4032	0.4023	0.4015	0.4007	0.3998	0.3990	0.3982	0.3973	0.3965
0.66	0.3957	0.3948	0.3940	0.3931	0.3922	0.3914	0.3905	0.3897	0.3888	0.3880
0.67	0.3871	0.3862	0.3854	0.3845	0.3836	0.3828	0.3819	0.3810	0.3801	0.3792
0.68	0.3784	0.3775	0.3766	0.3757	0.3748	0.3739	0.3730	0.3721	0.3712	0.3703
0.69	0.3694	0.3685	0.3676	0.3666	0.3657	0.3648	0.3639	0.3630	0.3621	0.3611
0.70	0.3602	0.3593	0.3583	0.3574	0.3565	0.3555	0.3546	0.3536	0.3527	0.3518
0.71	0.3508	0.3499	0.3489	0.3480	0.3470	0.3461	0.3451	0.3441	0.3432	0.3422
0.72	0.3412	0.3403	0.3393	0.3383	0.3373	0.3364	0.3354	0.3344	0.3334	0.3324
0.73	0.3314	0.3304	0.3295	0.3285	0.3275	0.3265	0.3255	0.3245	0.3235	0.3225
0.74	0.3215	0.3204	0.3194	0.3184	0.3174	0.3164	0.3154	0.3144	0.3133	0.3123
0.75	0.3113	0.3103	0.3092	0.3082	0.3071	0.3061	0.3051	0.3040	0.3030	0.3019
0.76	0.3009	0.2999	0.2988	0.2978	0.2967	0.2956	0.2946	0.2935	0.2925	0.2914
0.77	0.2903	0.2893	0.2882	0.2871	0.2861	0.2850	0.2839	0.2828	0.2818	0.2807
0.78	0.2796	0.2785	0.2774	0.2763	0.2753	0.2741	0.2731	0.2720	0.2709	0.2698
0.79	0.2687	0.2676	0.2664	0.2653	0.2642	0.2631	0.2620	0.2609	0.2598	0.2587
0.80	0.2575	0.2564	0.2553	0.2542	0.2531	0.2519	0.2508	0.2497	0.2485	0.2474
0.81	0.2462	0.2451	0.2440	0.2428	0.2417	0.2405	0.2394	0.2382	0.2371	0.2359
0.82	0.2348	0.2336	0.2324	0.2313	0.2301	0.2290	0.2278	0.2266	0.2255	0.2243
0.83	0.2231	0.2220	0.2208	0.2196	0.2184	0.2172	0.2160	0.2149	0.2137	0.2125
0.84	0.2113	0.2101	0.2089	0.2077	0.2065	0.2053	0.2041	0.2029	0.2017	0.2005
0.85	0.1993	0.1981	0.1969	0.1957	0.1944	0.1932	0.1920	0.1908	0.1896	0.1884
0.86	0.1871	0.1859	0.1847	0.1834	0.1822	0.1810	0.1797	0.1785	0.1773	0.1760
0.87	0.1748	0.1735	0.1723	0.1711	0.1698	0.1686	0.1673	0.1661	0.1648	0.1635
0.88	0.1623	0.1610	0.1598	0.1585	0.1572	0.1560	0.1547	0.1534	0.1522	0.1509
0.89	0.1496	0.1484	0.1471	0.1458	0.1445	0.1432	0.1419	0.1407	0.1394	0.1381
0.90	0.1368	0.1355	0.1342	0.1329	0.1316	0.1303	0.1290	0.1277	0.1264	0.1251
0.91	0.1238	0.1225	0.1212	0.1199	0.1186	0.1173	0.1159	0.1146	0.1133	0.1120
0.92	0.1107	0.1094	0.1080	0.1067	0.1054	0.1040	0.1027	0.1014	0.1000	0.0987
0.93	0.0974	0.0960	0.0947	0.0933	0.0920	0.0907	0.0893	0.0880	0.0866	0.0853
0.94	0.0839	0.0826	0.0812	0.0798	0.0785	0.0771	0.0758	0.0744	0.0730	0.0717
0.95	0.0703	0.0689	0.0676	0.0662	0.0648	0.0634	0.0621	0.0607	0.0593	0.0579
0.96	0.0565	0.0552	0.0538	0.0524	0.0510	0.0496	0.0482	0.0468	0.0454	0.0440
0.97	0.0426	0.0412	0.0398	0.0384	0.0370	0.0356	0.0342	0.0328	0.0314	0.0300
0.98	0.0286	0.0271	0.0257	0.0243	0.0230	0.0214	0.0201	0.0186	0.0172	0.0158
0.99	0.0140	0.0129	0.0115	0.0101	0.0086	0.0072	0.0058	0.0043	0.0029	0.0014

Appendix IV

SHORT TABLE OF THE FUNCTION, $h(p) = -p \log p - (1 - p) \log (1 - p)$

p	$h(p)$	p	$h(p)$
0.005	0.045415	0.130	0.557438
0.010	0.080793	0.135	0.570993
0.015	0.112364	0.140	0.584239
0.020	0.141441	0.145	0.597185
0.025	0.168661	0.150	0.609840
0.030	0.194392	0.155	0.622213
0.035	0.218878	0.160	0.634310
0.040	0.242292	0.165	0.646138
0.045	0.264765	0.170	0.657705
0.050	0.286397	0.175	0.669016
0.055	0.307268	0.180	0.680077
0.060	0.327445	0.185	0.690894
0.065	0.346981	0.190	0.701471
0.070	0.365924	0.195	0.711815
0.075	0.384312	0.200	0.721928
0.080	0.402179	0.205	0.731816
0.085	0.419556	0.210	0.741483
0.090	0.436470	0.215	0.750932
0.095	0.452943	0.220	0.760167
0.100	0.468996	0.225	0.769193
0.105	0.484648	0.230	0.778011
0.110	0.499916	0.235	0.786626
0.115	0.514816	0.240	0.795040
0.120	0.529361	0.245	0.803257
0.125	0.543564	0.250	0.811278

Appendix IV Table (contd. on page 396)

Appendix IV Table (contd. from page 395)

p	$h(p)$	p	$h(p)$
0.255	0.819107	0.380	0.958042
0.260	0.826746	0.385	0.961497
0.265	0.834198	0.390	0.964800
0.270	0.841465	0.395	0.967951
0.275	0.848548	0.400	0.970951
0.280	0.855441	0.405	0.973800
0.285	0.862175	0.410	0.976550
0.290	0.868721	0.415	0.979051
0.295	0.875093	0.420	0.981454
0.300	0.881291	0.425	0.983708
0.305	0.887317	0.430	0.985815
0.310	0.893178	0.435	0.987775
0.315	0.898861	0.440	0.989588
0.320	0.904381	0.445	0.991254
0.325	0.909736	0.450	0.992774
0.330	0.914925	0.455	0.994149
0.335	0.919953	0.460	0.995378
0.340	0.924819	0.465	0.996462
0.345	0.929523	0.470	0.997402
0.350	0.934068	0.475	0.998196
0.355	0.938454	0.480	0.998846
0.360	0.942683	0.485	0.999351
0.365	0.946755	0.490	0.999711
0.370	0.950672	0.495	0.999928
0.375	0.954434	0.500	0.100000

References

(Items marked with an asterisk * are in Russian. However, cover to cover English translations of many Russian Journals, e. g. *Problemy Peredachi Inform.*, *Uspekhi Mat. Nauk*, *Teoriya Veroyatn. i ee Primen.*, etc., have been published lately)

I—General

1. ABRAMSON, N. M. (1963). *Information Theory and Coding*. McGraw-Hill, New York.
2. ASH R. B. (1965). *Information Theory*. Interscience, New York.
3. ASHBY, W. R. (1965). *An Introduction to Cybernetics*. Chapman and Hall, London.
4. BAR-HILLEL, Y. and CARNAP, R. (1953). Semantic information. *Brit. Journal Phil. Sci.* 4, No. 14, 147-153; also in: Jackson, W. J. (ed.), *Communication Theory*, pp. 503-512 (1953). Butterworth, London and Academic Press, New York.
5. BRILLOUIN, L. (1967). *Science and Information Theory*, 2nd ed. Academic Press, New York.
6. CHERRY, C. (1966). *On Human Communication*, 2nd ed. MIT Press, Cambridge, Mass. and Wiley, New York.
7. CULLMAN, G. and DENIS-PAPIN, M. (1966). *Exercices de calcul informationnel avec leurs solutions*. Michel, Paris.
8. FANO, R. M. (1961). *Transmission of Information: A Statistical Theory of Communications*. MIT Press, Cambridge, Mass. and Wiley, New York.
9. FEINSTEIN, A. (1958). *Foundations of Information Theory*. McGraw-Hill, New York.
10. FEY, P. (1968). *Informationstheorie*. Academic-Verlag, Berlin.
11. GALLAGER, R. G. (1968). *Information Theory and Reliable Communication*. Wiley, New York.
- *12. GEL'FAND, I. M., KOLMOGOROV, A. N. and YAGLOM, A. M. On general definition of the amount of information. *Dokl. Akad. Nauk SSSR* III, No. 4, 745-748 (1956); Amount of information and entropy of continuous distributions. *Transactions of Third All-Union Mathematical Congress*, Volume 3, pp. 300-320. Izd. Akad. Nauk SSSR, Moscow (1958); GEL'FAND, I. M. and YAGLOM, A. M. (1957). Computation of the amount of information about a random function contained in another such function. *Uspekhi Mat. Nauk* 12, No. 1, 3-52. (English translation: *Amer Math. Soc. Translations*, ser 2, 12, 199-246 (1959).)
13. HINTIKKA, J. and SUPPES, P. (ed.) (1970). *Information and Inference*. Reidel, Dordrecht.
- *14. KOLMOGOROV, A. N. (1957). Theory of Transmission of Information, in: *Conference of Academy of Sciences of USSR on Scientific Problems of Automation, October 15-20, 1956: Plenary Session*. Izd. Akad. Nauk SSSR, Moscow. (English translation: *Amer. Math. Soc. Translations*, ser. 2, 33, 291-321.)

- *15. KOLMOGOROV, A. N. Three approaches to the quantitative definition of information. *Probl. Peredachi Inform.* 1, No. 1, 3-11 (1965); Logical basis for information theory and probability theory. *Probl. Peredachi Inform.* 5, No. 3, 3-7 and *IEEE Trans. Inform. Theory*, IT-14, 662-664 (1969).
- 16. MOLES, A. (1966). *Information Theory and Aesthetic Perception*. Translated by Cohen, J. E. Univ. of Illinois Press, Urbana, Ill.
- 17. PIERCE, J. R. (1961). *Symbols, Signals and Noise*. Harper, New York.
- *18. POLETAYEV, I. A. (1958). *The Signal*. Soviet Radio, Moscow.
- 19. QUASTLER, H. (ed.) (1955). *Information Theory in Psychology*. Free Press of Glencoe, Glencoe, Ill.
- *20. SCHREIDER, YU. A. On a model of semantic theory of information. *Problemy Kibernetiki* 13, 233-240 (1965). On the semantic aspects of information theory. In: *Information and Cybernetics*. Soviet Radio, Moscow (1967).
- 21. SHANNON, C. E. (1948). A mathematical theory of communication. *Bell System Tech. J.* 27, 379-423, 623-656. Reprinted in: Shannon, C. E. and Weaver, W. (1949). *The Mathematical Theory of Communication*. Univ. of Illinois Press, Urbana, Ill.
- 22. WIENER, N. (1961). *Cybernetics, or Control and Communication in the Animal and the Machine*, 2nd ed. MIT Technology Press, Cambridge, Mass.
- 23. WOLFOWITZ, J. (1964). *Coding Theorems of Information Theory*, 2nd ed. Springer Verlag, Berlin.
- 24. WOODWARD, P. M. (1953). *Probability and Information Theory with Application to Radar*. Pergamon Press, London.
- 25. WOZENCRAFT, J. M. and JACOBS, I. M. (1965). *Principles of Communication Engineering*. Wiley, New York.
- 26. YOCKEY, H. P., PLATZMAN, R. L. and QUASTLER, H. (eds.) (1958). *Symposium on Information Theory in Biology*. Pergamon Press, London and New York.
- *27. ZVONKIN, A. K. and LEVIN, L. A. (1970). The complexity of finite objects and the foundation of the concepts of information and randomness based on algorithm theory. *Uspekhi Mat. Nauk (Soviet Math. Surveys)* 25, No. 6, 85-127.

II—Specific

Chapter 1

- 28. CULLBERTSON, Y. T. (1958). *Mathematics and Logic for Digital Devices*. Van-Nostrand, Princeton.
- 29. DIAMOND, S. (1964). *The World of Probability*. Basic Books, New York.
- 30. FELLER, W. (1968). *An Introduction to Probability Theory and its Applications*, Volume 1. Wiley, New York.
- 31. GNEDENKO, B. V. and KHINCHIN, A. YA. (1961). *An Elementary Introduction to the Theory of Probability*. Dover, New York.
- 32. HODGES, S. L. and LEHMANN, E. L. (1965). *Elements of Finite Probability*. Holden Day, San Francisco.
- 33. KAC, M. (1964). Probability theory. In: *Mathematics in the Modern World* (Intro. by Morris Kline). Freeman, New York.
- 34. KEMENY, J. G., MIRKIL, H., SNELL, J. L. and THOMPSON, G. L. (1959). *Finite Mathematical Structures*. Prentice-Hall, Englewood Cliffs, New Jersey.
- 35. KOLMOGOROV, A. N. (1964). Probability Theory. In: *Mathematics: its Contents, Methods and Meaning*, Volume 2. English translation from Russian. MIT Press, Cambridge, Mass.

36. MESHALKIN, L. D. (1972). *Collection of Problems in Probability Theory*. English translation from Russian. Noordoff, Leyden.
37. MOSTELLER, F. (1965). *Fifty Challenging Problems in Probability*. Addison-Wesley, Reading, Mass.
38. MOSTELLER, F., ROURKE, R. E. K. and THOMAS, G. B. (1961). *Probability with Statistical Applications*. Addison-Wesley, Reading, Mass.
39. NEYMAN, J. (1951). *First Course in Probability and Statistics*. Holt, Rinehart and Winston, New York.
40. YAGLOM, A. M. and YAGLOM, I. M. (1964). *Challenging Mathematical Problems with Elementary Solutions*, Volume 1. English translation from Russian. Holden-Day, San Francisco.

Chapter 2

41. ACZÉL, Y., FORTE, B. and NG, C. T. (1974). Why the Shannon and Hartley entropies are 'natural'? *Advances Appl. Probab.* 6, No. 1, 131-146.
42. ATTNEAVE, F. (1959). *Applications of Information Theory to Psychology: A Summary of Basic Concepts, Methods and Results*. Holt-Dryden, New York.
43. DARÓCZY, Z. (1970). Generalized information functions. *Information and Control* 16, No. 1, 36-51.
- *44. DOBRUSHIN, R. L. (1958). Transmission of information through channels with feedback. *Teoriya Veroyatn. i ee Priman. (Theory of Prob. and Appl.)* 3, No. 4, 395-412.
- *45. FADDEEV, D. K. (1956). On the concept of entropy of a finite probabilistic scheme. *Uspekhi Mat. Nauk (Soviet Math. Surveys)* 11, No. 1 (67), 227-231.
46. HICK, W. E. (1952). On the rate of gain of information. *Quart. J. Experimental Psychology* 4, No. 1, 11-26.
47. HYMAN, R. (1953). Stimulus information as a determinant of reaction times. *Journ. of Experimental Psychology* 45, No. 3, 188-196.
48. LEONARD, J. A. (1961). Choice reaction time experiments and information theory. In: C. Cherry (ed.), *Information Theory*, pp. 137-146. Butterworth, London and Academic Press, New York. See also LUCE, R. D. (1960). The theory of selective information and some of its behavioural applications. In: *Developments in Mathematical Psychology, Information, Learning and Tracking*, pp. 5-119. Free Press of Glencoe, Glencoe, Ill.
- *49. LEONTIEV, A. N. and KRINCHIK, E. P. (1961). On applications of information theory to specific psychological investigations. *Voprosy psichol.* No. 4, 25-46.
- *50. LOMOV, B. F. (1966). *Man and Technology (An Outline of Engineering Psychology)*. Soviet Radio, Moscow.
- *51. NIKOLAEV, V. I. (1965). The determination of time needed by an operator to solve problems of ship engine control. *Izv. Akad. Nauk SSSR (Ser. Energetika and Transport)* No. 4, 130-145.
52. WELFORD, A. T. (1960). The measurement of sensory-motor performance: Survey and reappraisal of twelve years progress. *Ergonomics* 3, No. 3, 189-231.

Chapter 3

53. BELLMAN, R. and GLUSS, B. (1961). On various versions of the defective coin problem. *Information and Control* 4, Nos. 2-3, 118-131, errata on p. 391, same volume, No. 4.
54. DEVIDÉ, V. (1959). Ein Problem über Wagen. *Elemente der Math.* 10, No. 1, 11-15.
55. FORD, L. R. and JOHNSON, S. M. (1959). A tournament problem. *American Math. Monthly* 66, No. 5, 387-389.
56. KELLOGG, P. J. and KELLOGG, D. J. (1954). Entropy of information and the odd ball problem. *Journ. of Appl. Phys.* 25, No. 11, 1438-1439.

- *57. KISLITSIN, S. S. (1962). Present status of the search theory. *Uspekhi. Mat. Nauk* 17, No. 1, 243-244.
- *58. KISLITSIN, S. S. (1963). The refinement of an estimate of the least mean number of comparisons, which are necessary for well ordering of a finite collection. *Vestnik Leningrad Gos. Univ.* 19, No. 4, 143-145.
- *59. KORDEMSKII, B. A. (1965). *Mathematical Sharpness*. Nauka, Moscow.
- *60. PARKHOMENKO, P. P. (1970). Theory of questionnaires : A review. *Automatika i Telemekhanika*, No. 4, 140-159.
- 61. PICARD, C.-F. *Theorie des questionnaires*. Gauthier-Villars, Paris (1965); *Graphes et questionnaires*. Tome II. *Questionnaires*. Gauthier-Villars, Paris (1972).
- 62. SHKLARSKY, D. O., CHENTSOV, N. N. and YAGLOM, I. M. (1962). *The USSR Olympiad Problem Book*. English translation from Russian. Freeman, San Francisco.
- *63. STEINHAUS, H. (1959). *Hundred Problems*. Fizmatgiz, Moscow.

Chapter 4, Section 1

- 64. GILBERT, E. N. and MOORE, E. F. (1959). Variable length binary encodings. *Bell System Tech. J.* 38, 933-967.
- 65. SARDINAS, A. A. and PATTERSON, G. W. *A Necessary and Sufficient Condition for the Unique Decomposition of Coded Messages*. Research Division Report 50-57 (1950), Moore School of Electrical Engineering, University of Pennsylvania, Philadelphia, Pa; also, *IRE Convention Record*, Part 8, pp. 104-108 (1953).

Chapter 4, Section 2

- 66. HUFFMAN, D. A. (1952). A method for the construction of minimum redundancy codes. *Proc. IRE* 40, No. 10, 1098-1101.
- 67. KARUSH, J. L. (1961). A simple proof of an inequality of McMillan. *IRE Trans. Inform. Theory* IT-7, No. 2, 118.
- 68. MCMILLAN, B. (1953). The basic theorems of information theory. *Ann. Math. Stat.* 24, No. 2, 196-219.
- 69. MCMILLAN, B. (1956). Two inequalities implied by unique decipherability. *IRE Trans. Inform. Theory* IT2, 115-116.

Chapter 4, Section 3

- *70. ALEKSEEV, P. M. (1971). Frequencies count dictionaries of English and their practical applications. In: *Statistics of Speech and Automatic Analysis of Texts*, pp. 160-170. Nauka, Leningrad.
- 71. APOSTEL, L., MANDELBROT, B. and MORE, A. (1957). *Logique, language et théorie de l'information*. Presses Universitaires de France, Paris.
- 72. BARELL, B. C., AIR, G. M. and HUTCHISON, C. A. (1976). Overlapping genes in bacteriophage X 174. *Nature*, 264, No. 5581, 34-41.
- 73. BARNARD, G. A. (1955). Statistical calculation of word entropies for four western languages. *IRE Trans. Information. Theory* IT-1, No. 1, 49-53.
- *74. BASHARIN, G. P. (1959). On statistical estimate for the entropy of a sequence of independent random variables. *Teoriya Veroyatn. i ee Primen. (Theory of Prob. and Appl.)* 4, No. 3, 361-364.
- 75. BELVITCH, V. (1956). Théorie de l'information et statistique linguistique. *Bulletin Acad. Royale Belgique (Classe de sciences)*, 419-436.

76. BERRY, J. (1953). Some statistical aspects of conversational speed. In: Jackson, W. (ed.), *Communication Theory*, pp. 392-401. Butterworths, London, and Academic Press, New York.
77. BLACK, J. W. (1954). The information of sounds and phonetic digrams of one- and two-syllable words. *Journ. Speech Hearing Disorders* 19, 397-411; DENES, P. (1963). On the statistics of spoken English. *Journ. Acoust. Soc. Amer.* 35, No. 6, 892-904.
78. BLUHME, H. (1963). Three-dimensional crossword puzzles in Hebrew. *Information and Control* 6, No. 3, 306-309.
79. BLYTH, C.R. (1958). *Note on Estimating Information*. Techn. Report. No. 17, Dept of Statistics, Stanford University.
80. BROOKS, F. P., HOPKINS, A.L., NEUMANN, P. G. and WRIGHT, W. V. (1975). An experiment in musical composition. *IRE Trans. Electron. Comput.* EC-6, No. 3, 175-182.
81. BURTON, N. G. and LICKLIDER, J. C. R. (1955). Long-range constraints in the statistical structure of printed English. *Amer. Journ. of Psychology* 68, No. 4, 650-653.
82. CARSON, D. H. (1961). Letter constraints within words in printed English. *Kybernetik* 1, 46-54.
83. CARTERETTE, E. C. and JONES, M. H. (1963). Redundancy in children's texts. *Science* 140, No. 3573, 1309-1311.
84. CHERRY, E. C., HALLE, M. and JAKOBSON, R. (1953). Towards the logical description of languages in their phonemic aspect. *Language* 29, No. 1, 34-46.
85. COHEN, J. E. (1962). Information theory and music. *Behav. Sc.* 7, No. 2, 137-163.
86. COVER, T. M. and KING, R. C. (1976). *A Convergent Gambling Estimate of the Entropy of English*. Tech. Report No. 22, Dept. of Statistics, Stanford University. See also *IEEE Trans. Inform. Theory* IT-24, No. 4, 413-421 (1978).
87. CRICK, F. H. C. (1962). The genetic code. *Sci. Amer.* 207, No. 4, 66-74; NIRENBERG, M. W. (1963). The genetic code: II. *Sci., Amer.* 208, No. 3, 80-94; CRICK, F. H. C. (1966). The genetic code III. *Sci. Amer.* 215, No. 4, 55-61.
88. CRICK, F. H. C., GRIFFITH, J. S. and ORGEL, L. E. (1957). Codes without commas. *Proc. Nat. Acad. Sci. USA* 43, 416-421.
89. DEUTSCH, S. (1957). A note on some statistics concerning typewritten or printed material. *IRE Trans. Inform. Theory* IT-3, No. 2, 136-147.
90. DEWEY, G. (1923). *Relative Frequency of English Speech Sounds*. Harvard University Press, Cambridge (Mass.).
- *91. DOBRUSHIN, R. L. (1961). Mathematical methods in linguistics. *Matematicheskoe Prosveshcheniye (Mathematical Education; new series)* 6, 37-60. Fizmatgiz, Moscow.
92. ELDRIDGE, R. C. (1911). *Six Thousand Common English Words*. Niagara Falls, N. Y.
93. ENDRES, W. (1973). A comparison of the redundancy in the written and spoken language. In: Petrov, B. N. and Csáki, F. (eds.), *Proceedings of Second International Symposium on Information Theory, Tsakhadzor, Armen. SSR*, pp. 53-59. Akadémiai Kiadó, Budapest.
94. Fast Data Communication (1963). *Sci. News Letters* 83, No. 1, 5.
95. FOY, W. H. (1964). Entropy of simple line drawings. *IEEE Trans. Inform. Theory* IT-10, No. 2, 165-167.
96. FRADIS, A., MIHAILESCU, L. and VOINESCU, I. (1967). L'entropie et l'énergie informationnelle de la langue roumaine parlée. *Revue roumaine de linguistique* 12, No. 4, 331-339.
97. FRICK, F. C. and SUMBY, W. H. (1952). Control tower language. *Journ. Acoust. Soc. Amer.* 24, 595-596.
98. FRITZ, E. L. and GRIER, G. W. (1955). Pragmatic communication; a study of information flow in air traffic control. In: Quastler, H. (ed.), *Information Theory in Psychology*. The Free Press of Glencoe, Glencoe, Ill.
- *99. FROLUSHKIN, V. G. (1959). Analysis of statistical structure of phototelegram texts. *Elektrosvyaz (Electrical Communication)*, No. 5, 65-70.

100. GAMOW, G. (1954). Possible relation between deoxyribonucleic acid, and protein structures. *Nature* **173**, 318.
101. GAMOW, G., RICH, A. and YČAS, M. (1956). The problem of information transfer from the nucleic acids to proteins. *Advances Biol. Medical Phys.* **4**, 23-68; GAMOW, G. and YČAS, M. (1958). Cryptographic approach to protein synthesis. YČAS, M. (1958). The protein text. Both in: Yockey, H. P., Platzman, R. L. and Quastler, H. (eds.), *Symposium on Information Theory in Biology*. Pergamon Press, London and New York.
102. GAMOW, G. and YČAS, M. (1955). Statistical correlation of protein and ribonucleic acid composition. *Proc. Nat. Acad. Sci. USA* **41**, 1011-1019.
- *103. GARMASH, V. A. and KIRILLOV, N. E. (1959). An experimental investigation of the statistics of phototelegraphic messages. *Nauchn. Dokl. Vyssh. Shkoly, Ser. Radiotekhnika i Elektr. (Sci. Rep. Colleges and Univ. USSR, Ser. Radioengineering and Electronics)*. No. 1, 37-42.
104. GOLOMB, S. W., WELCH, L. R. and DELBRÜCK, M. (1958). Construction and properties of comma-free codes. *Kgl. Danske Vid. Selsk. Biol. Medd.* **23**, No. 9, 1-34.
105. GRIGNETTI, M. (1964). A note on the entropy of words in printed English. *Information and Control* **7**, No. 3, 304-306.
106. HATON, J. P. and LAMOTTE, M. (1971). Étude statistique des phonèmes et diphonèmes dans le français parlé. *Revue d'acoustique* **4**, No. 16, 258-262.
107. HILLER, L. and BEAUCHAMP, J. (1965). Research in music with electronics. *Science* **150**, No. 3693, 161-169.
- *108. IBRAGIMOV, T. I. (1969). Researches in syllabic organization of the Tatar language. *Uchen. Zap. Kazan. Gos. Univ. (Scient. Trans. of Kazan State Univ.)* **129**, No. 4, 101-108.
109. JACOBSON, H. (1951). The informational capacity of the human eye. *Science* **113**, No. 2933, 292-293.
110. JACOBSON, H. (1951). Information and the human ear. *Journ. Acoust. Soc. Amer.* **23**, No. 4, 463-471.
111. JAMISON, D. and JAMISON, K. (1968). A note on the entropy of partially known languages. *Information and Control* **12**, No. 2, 164-167.
112. KAYSER, G. A. (1960). Zur Entropie schreibmaschinengeschriebener Textvorlagen. *Nachrichtentechn. Z.* **13**, No. 5, 219-224.
- *113. KAZARYAN, R. A. (1961). The evaluation of the entropy of Armenian language. *Izv. Akad. Nauk Armen. SSR (Physical and mathematical sciences)* **14**, No. 4, 161-173; LENSKIĬ, D. N. (1962). On estimating entropy of printed texts in Adygei language. *Uchen. Zap. (Sci. Reports) of Kabardino-Balkar Univ. (Physico-mathematical series)* No. 16, 165-166; IBRAGIMOV, T. I. (1964). An evaluation of the interrelation of letters in the Tatar literary language. *Scientific Trans. of Kazan State Univ.* **124**, No. 2, 141-145.
114. KELLY, D. H. (1962). Information capacity of a single retinal channel. *IRE Trans. Inform. Theory* **IT-8**, No. 3, 221-226.
- *115. KHARKEVICH, A. A. (1955). *Outline of the General Theory of Communications*. Gostekhizdat, Moscow.
- *116. KONDRATOV, A. M. (1963). Information theory and prosody (entropy of the rhythm of Russian speech). *Problemy Kibernetiki* **9**, 279-286.
117. KREUZER, H. and GUNZENHÄUSER, R. (eds.) (1965). *Mathematik und Dichtung*. Nymphenburger Verlagshandlung, München.
118. KÜPFMÜLLER, K. (1954). Die Entropie der deutschen Sprache. *Fernmeldetechn. Z.*, No. 6, 265-272.
119. KÜPFMÜLLER, K. (1959). Informationsverarbeitung durch den Menschen. *Nachrichtentechn. Z.*, No. 2, 68-74.
- *120. LEBEDEV, D. S. and GARMASH, V. A. (1958). On the possibility of increase in transmission rate of telegraphic messages. *Elektrosvyaz (Electrical Communication)*, No. 1, 58-69.

- *121. LEBEDEV, D. S. and PIL, E. I. (1959). Experimental researches in statistics of television messages. *Tekhnika Kino i Televiz. (Film and Television Technique)*, No. 3, 37-39.
- *122. LEBEDEV, D. S. and TSUKKERMAN, I. I. (1965). *Television and Information Theory*. Energiya, Moscow.
- 123. LIMB, J. O. (1968). Entropy of quantized television signals. *Proc. Inst. Elec. Engg. (Proc. IEE)* 115, No. 1, 16-20.
- 124. MAIXNER, V. (1971). Some remarks on entropy prediction of natural language texts. *Information Stor. Retr.* 7, 293-295.
- 125. MANDELBROT, B. (1953). An informational theory of the statistical structure of language. In: Jackson, W. (ed.), *Communication Theory*. Butterworths, London, and Academic Press, New York.
- *126. MANDELBROT, B. (1957). A law of Berry and definition of "stress". In: *Theory of Transmission of Information*, pp. 248-254, Izd. Inostr. Liter., Moscow. See also MANDELBROT, B. (1957). A note on a law of Berry and on insistence stress. *Information and Control* 1, No. 1, 76-81.
- 127. MANDELBROT, B. (1977). *Fractals. Form, Chance and Dimension*. W.H. Freeman, San Francisco.
- 128. MANFRINO, R. L. L'entropia della lingua italiana ed il suo calcolo. *Alta frequenza* 29, No. 1, 4-29 (1960); Printed Portuguese (Brazilian) entropy statistical calculation. *IEEE Trans. Inform. Theory* IT-16, No. 1, 122 (1970); HANSSON, H. (1960). The entropy of the Swedish language. *Trans. of the Second Prague Conference on Information Theory, Statistical Decision Functions and Random Processes*, pp. 215-270; DOLEZEL, L. (1963). Předběžný odhad entropie a redundance psané češtiny. *Solvo a Slovesnost* 24, No. 3, 165-175; ZITEK, F. (1964). Quelques remarques au sujet de l'entropie du tchèque. *Trans. of the Third Prague Conference on Information Theory, Statistical Decision Functions and Random Processes*, pp. 841-846; NICOLAU, E., SALA, C. and ROCERIC, A. (1959). Observatii asupra entropiei limbii romane. *Studii cercetari lingvist.* 10, No. 1, 35-54; WANAS, M. A., ZAYED, A. I., SHAKER, M. M. and TAHA, E. H. (1976). First-, second- and third-order entropies of Arabic text. *IEEE Trans. Inform. Theory* IT-22, No. 1, 123.
- 129. MARCUS, S. (1967). Entropie et énergie poétique, *Cahiers de linguistique théorique et appliquée* 4, 171-180.
- 130. MICHEL, S. W. (1958). Statistical encoding for text and picture communications. *Commun. and Electr.*, 35, 33-36.
- 131. MILLER, G. A. (1951). Speech and language. In: Stevens, S. S. (ed.), *Handbook of Experimental Psychology*. Wiley, New York.
- 132. NEIDHARDT, P. *Einführung in die Informationstheorie*. VEB Verlag Technik, Berlin (1957); *Informationstheorie und automatische Informationsverarbeitung*. VEB Verlag Technik, Berlin (1964).
- 133. NEMETZ, T. (1972). On the experimental determination of the entropy. *Kybernetik* 10, 137-139.
- 134. NEMETZ, T. and SIMON, J. (1978). On estimating the entropy of written Hungarian by gambling technique. Submitted to the *Trans. of the Eighth Prague Conference on Inform. Theory, Statistical Decision Functions and Random Processes* (Prague, August 28-September 1, 1978).
- *135. NEVEL'SKII, P. B. and ROSENBAUM, M. D. (1971). Guessing professional text among specialists and non-specialists. In: *Statistics of Speech and Automatic Analysis of Texts*, pp. 134-148. Nauka, Leningrad.
- 136. NEWMAN, E. B. (1959). Men and information : a psychologist's view. *Nuovo Cimento Suppl.* 13, No. 2, 539-559.
- 137. NEWMAN, E. B. and GERSTMAN, L. J. (1952). A new method for analyzing printed English. *Journ. of. Experimental Psychology* 44, No. 2, 114-125.

138. NEWMAN, E. B. and WAUGH, N. C. (1960). The redundancy of texts in three languages. *Information and Control* 3, No. 2, 141-153.
139. OLSON, H. and BELAR, H. (1961). Aid to music composition employing a random probability system. *Journ. Acoust. Soc. America* 33, No. 9, 1163-1170.
140. PAISLEY, W. J. (1966). The effects of authorship, topic, structure and time of composition on letter redundancy in English texts. *Journ. Verbal Learning and Verbal Behaviour* 5, No. 1, 28-34.
141. PARKS, J. R. (1965). Prediction and entropy of half-tone pictures. *Behavioral Sci.* 10, 436-445.
- *142. PESHKOVSKII, A. M. (1925). Ten thousand sounds. In: *Collection of Papers*, pp. 167-191. Gosud. Izdatelstvo, Leningrad-Moscow.
- *143. PETROVA, N. V. (1965). The entropy of printed French. *Izv. Akad. Nauk SSSR (Ser. of literature and language)* 24, No. 1 63-67. See also PETROVA, N. V., PIOTROVSKII, R. G. and GIRAUD, R. (1964). L'entropie du français écrit. *Bull. Soc. de linguistique de Paris* 58, No. 1, 130-152.
144. PFAFFELHUBER, E. (1971). Error estimation for the determination of entropy and information rate from relative frequencies. *Kybernetik* 8, 50-51.
145. PINKERTON, R. C. (1956). Information theory and melody. *Scient. Amer.* 194, No. 2, 77-86.
- *146. PIOTROVSKAYA, A. A., PIOTROVSKII, R. G. and RAZZHIVIN, K. A. (1961). The entropy of the Russian language. *Voprosy yazykoznaviya (Problems of Linguistics)*, No. 6, 115-130.
- *147. PIOTROVSKII, R. G. (1968). *Information Measurements of Language*. Nauka, Leningrad.
- *148. PIOTROVSKII R. G., BEKTAEV, K. B. and PIOTROVSKAYA, A. A. (1977). *Mathematical Linguistics*. Vysshaya Shkola (Publ. House 'Higher School'), Moscow.
149. PRATT, F. (1942). *Secret and Urgent*. Doubleday, Garden City (N. Y.).
150. QUASTLER, H. (1956). Studies of human channel capacity. In: Cherry, C. (ed.), *Information Theory, Third London Symposium*, pp. 361-371. Butterworths, London.
151. RAMAKRISHNA, B. S. and SUBRAMANIAN, R. (1958). Relative efficiency of English and German languages for communication of semantic content. *IRE Trans. Inform. Theory* IT-4, No. 3, 127-129.
152. REZA, F. M. (1961). *An Introduction to Information Theory*. McGraw-Hill, New York.
153. ROLAND, M. (1967). Die Entropieabnahme bei Abhängigkeit zwischen mehreren simultanen Informationsquellen und bei Übergang zu Markoff-Ketten höherer Ordnung, untersucht an musikalischen Beispielen. *Forschungsber. Landes Nordrhein-Westfalen*, No. 1768, pp. 39, 41, 43-44, 79-80.
- *154. RYCHKOVA, N. (1961). Linguistics and mathematics. *Nauka i Zhizn (Science and Life)*, No. 9, 76-77.
- *155. SAVCHUK, A. P. (1964). On the evaluation of the entropy of language using the method due to Shannon. *Teoriya Veroyatn. i ee Primen. (Theory of Prob. and Appl.)* 9, No. 1, 154-157.
156. SCHÖBER, H. (1957). Grundlegende Bemerkungen zur Anwendbarkeit der Informationstheorie auf die Optik. *Wiss. Zeitschr. Hochschule Elektrotechn. Ilmenau* 3, No. 3-4, 273-276.
157. SCHREIBER, W. F. (1956). The measurement of third order probability distributions of television signals. *IRE Trans. Inform. Theory* IT-2, No. 3, 94-105.
158. SHANNON, C. E. (1949). Communication in the presence of noise. *Proc. IRE* 37, No. 1, 10-21.
159. SHANNON, C. E. (1951). Prediction and entropy of printed English. *Bell System Tech. Journ.* 30, No. 1, 50-64.
160. SIROMONEY, G. (1963). Entropy of Tamil prose. *Information and Control* 6, No. 3, 297-300; RAJAGOPALAN, K. R. (1965). A note on entropy of Kannada prose. *Information*

- and *Control* 8, No. 6, 640-644; BALASUBRAHMANYAM, P. and SIROMONEY, G. (1968). A note on entropy of Telugu prose. *Information and Control* 13, No. 4, 281-285; RAMAKRISHNA, B. S., NAIR K. K., CHIPLUNKAR, V. N., ATAL, B. S., RAMACHANDRAN, V. and SUBRAMANIAN, R. (1961). Relative efficiencies of Indian languages. *Nature* 189, No. 4768, 614-617.
161. SIROMONEY, G. (1964). An information-theoretical test for familiarity with a foreign language. *Journ. Psychol. Researches* 8, 1-6.
162. SIROMONEY, G. and RAJAGOPALAN, K. R. (1964). Style as information in Karnatic music. *Journ. Music Theory* 8, No. 2, 267-272.
- *163. SMIRNOV, O. L. and YEKIMOV, A. V. (1967). Entropy of the Russian telegraphic text. *Trudy Leningr. Inst. Aviatsion. Mashinostr. (Transactions of Leningrad Inst. of Mechan. Engg. for Aircraft Industry)*, No. 54, 76-84.
164. SZIKLAI, G. C. (1956). Some studies in the speed of visual perception. *IRE Trans. Inform. Theory* IT-2, No. 3, 125-128.
165. TARNÓCZY, T. (1961). A jeloszlás és a hírtartalom nyelveket meghatározó tulajdonságairól. *Nyelvtudományi Közlemények* 63, 161-178.
- *166. TEMNIKOV, F. E., AFONIN, V. A. and DMITRIEV, V. I. (1971). *Theoretical Basis of Information Techniques*. Energiya. Moscow.
167. THORNDIKE, E. L. (1932). *A Teacher's Word Book*. New York.
168. TSANNES, N. S., SPENCER, R. V. and KAPLAN, A. J. (1970). On estimating the entropy of random fields. *Information and Control* 16, No. 1, 1-6.
- *169. URBACH, V. YU. (1963). On taking note of the correlation between alphabet letters when amount of information in a message is evaluated. *Problemy Kibernetiki* 10, 111-117.
- *170. USPENSKI, V. A. (1964). A model for the notion of phoneme. *Voprosy yazykoznaniya (Problems of Linguistics)*, No. 6, 39-53.
- *171. VASILIEV, R. R. (1957). On statistical methods for transmission of phototelegrams. *Radiotekhnika i Elektronika (Radio Engineering and Electronics)* 2, No. 2, 136-143.
172. VOINESCU, I., FRADIS, A. and MIHAILESCU, L. The first degree entropy of phonemes in aphasics. *Revue roumaine de neurologie* 4, No. 1, 67-79 (1967); Second order entropy of phonemes and rank-frequency relation of biphonemic groups in aphasics. *Revue roumaine de neurologie*, 5, No. 2, 111-120 (1968); First order entropy of words in aphasics. *Cybernetica* 12, No. 1, 39-49 (1969). See also KREINDLER, A. and FRADIS, A. (1970). Theoria informatiei, limbajul si afazia. Chapter IX in : *Afazia*. Ed. Acad. Rep. Social Romania, Bucuresti.
173. WELTNER, K. (1973). *The Measurement of Verbal Information in Psychology and Education*. Springer-Verlag, Berlin-Heidelberg-New York.
- *174. YAGLOM, I. M., DOBRUSHIN, R. L. and YAGLOM, A. M. (1960). Information theory and linguistics. *Voprosy yazykoznaniya (Problems of Linguistics)*, No. 1, 100-110.
175. YCAS, M. (1969). *The Biological Code*. North-Holland, Amsterdam.
176. YOUNGBLOOD, J. E. (1958). Style as information. *Journ. Music Theory* 2, No. 1, p. 24 et seq.
- *177. ZARIPOV, R. Kh. (1971). *Cybernetics and Music*. Nauka, Moscow.
- *178. ZINDER, L. R. (1958). On the linguistic probability. In : *Statistical Problems of Speech*, pp. 58-61. Izd. Leningrad State Univ., Leningrad.
179. ZIFF, G. K. (1963). *Human Behaviour and the Principle of Least Effort*. Addison-Wesley, Cambridge, Mass.

Chapter 4: Section 4

180. BARNARD, G. A. (1956). Simple proofs of simple cases of the coding theorem. In: Cherry, C. (ed.), *Information Theory*, pp. 96-100. Butterworths, London.

- *181. DOBRUSHIN, R. L. (1962). Asymptotic estimates of the probability of error in the transmission of information through a discrete memoryless communication channel with a symmetric matrix of transition probabilities. *Teoriya Veroyatn. i ee Primen. (Theory of Prob. and Appl.)* 7, No. 3, 283-311.
- 182. ELIAS, P. (1956). Coding for two noisy channels. In: Cherry, C. (ed.), *Information Theory*. Butterworths, London.
- 183. GALLAGER, R. G. (1965). A simple derivation of coding theorem and some applications. *IEEE Trans. Inform. Theory* IT-11, No. 1, 3-18.
- 184. GILBERT, E. N. (1952). A comparison of signaling alphabets. *Bell System Tech. Journ.* 31, No. 3, 502-522.
- 185. SHANNON, C. E. (1956). The zero-error capacity of a noisy channel. *IRE Trans. Inform. Theory* IT-2, No. 1, 8-19.
- 186. SHANNON, C. E. (1957). Certain results in coding theory for noisy channels. *Information and Control* 1, No. 1, 6-25.
- 187. SLEPLIAN, D. (1959). Coding theory. *Nuovo Cimento Suppl.* 13, No. 2, 373-388.
- *188. ZAREMBA, S. K. Note on the fundamental theorem for a discrete noisy channel. In: *Theory of Transmission of Information*, pp. 28-31. Izd. Inostr. Liter., Moscow (1957). Cf. also ZAREMBA S. K. (1956). Discussion of the paper by G. A. Bernard. In: Cherry, C. (ed.), *Information Theory*, pp. 100-102. Butterworths, London, and Academic Press, New York.

Chapter 4. Section 5

- 189. AHLWEDE, R. (1971). Group codes do not achieve Shannon's channel capacity for general discrete channels. *Ann. Math. Stat.* 42, No. 1, 224-240.
- 190. BERLEKAMP, E. (1968). *Algebraic Coding Theory*. McGraw-Hill, New York.
- 191. BIRKHOFF, G. and BARTEE, T. C. (1970). *Modern Applied Algebra*. McGraw-Hill, New York.
- 192. BOSE, R. C. and RAY-CHAUDHURI, D. K. (1960). On a class of error correcting binary group codes. *Information and Control* 3, 68-79. Further results on error correcting binary group codes. *Information and Control* 3, 279-290.
- 193. CULLMANN, G. (1967). *Codes détecteurs et correcteurs d'erreurs*. Dunod, Paris.
- *194. DOBRUSHIN, R. L. (1963). Asymptotic optimality of group and systematic codes for some channels. *Teoriya Veroyatn. i ee Primen. (Theory of Prob. and Appl.)* 8, No. 1, 52-66.
- *195. DOBRUSHIN, R. L. (1966). Theory of optimal coding of information. In: Berg, A. I. (ed.), *Cybernetics must Serve the Communism* 3, 13-45.
- 196. DRYGAS, H. (1965). Verschlüsselungstheorie für symmetrische Kanäle. *Zeitschr. für Wahrscheinlichkeitstheorie and verw. Gebiete* 4, 121-143.
- 197. ELIAS, P. (1961). Coding and decoding. In: Baghdady, E. J. (ed.), *Lectures on Communication System Theory*. McGraw-Hill, New York.
- 198. FORNEY, G. D. (1966). *Concatenated Codes*. MIT Press, Cambridge, Mass.
- 199. GALLAGER, R. G. (1963). *Low-density Parity-Check Codes*. MIT Press, Cambridge, Mass.
- *200. GABIDULIN, E. M. (1967). Bounds to error probability for decoding by using memoryless linear codes. *Probl. Peredachi Inform.* 3, No. 2, 55-62.
- 201. GILBERT, W. J. (1976). *Modern Algebra with Applications*. Wiley, New York.
- 202. GORENSTEIN, D. C., PETERSON, W. W. and ZIERLER, N. (1960). Two-error correcting Bose-Chaudhuri codes are quasi-perfect. *Information and Control* 3, 291-294.
- 203. HAMMING, R. W. (1950). Error detecting and error correcting codes. *Bell System Tech. J.* 29, 147-160.
- 204. HOCQUENGHEM, A. (1959). Codes correcteurs d'erreurs. *Chiffres* 2, 147-156.

- *205. KASAMI, T., TOKURA, N., IVADARI, E. and INAGAKI, Y. (1978). *Coding Theory*. Russian translation from Japanese. MIR Publishers, Moscow.
- 206. KAUFMANN, A. (1968). *Introduction a la combinatoire en vue des applications*. Dunod, Paris.
- *207. KOLBSNIK, V. D. and MIRONCHIKOV, E. T. (1968). *Decoding of Cyclic Codes*. Svyaz (Communication), Moscow.
- 208. LEVINSON, N. (1970). Coding theory: a counter-example to G. H. Hardy's conception of applied mathematics. *Amer. Math. Monthly* 77, No. 3, 249-258.
- 209. LIN, S. (1970). *Introduction to Error-Correcting Codes*. Prentice-Hall, Englewood Cliffs (N. J.).
- 210. LINT, J. H. VAN. *Coding Theory*. Springer-Verlag, Berlin (1971); LINT, J. H. van. Non-existence theorems for perfect error-correcting codes. In: *Computing Algebra and Number Theory (Math. Symposia)* 4, 89-95. American Math. Soc., Providence, R. I. (1971).
- 211. MACWILLIAMS, F. J. and SLOANE, N. J. A. (1977). *The Theory of Error-Correcting Codes*, Parts I and II. North Holland, Amsterdam.
- 212. PETERSON, W. W. and WELDON, E. Y. (1972). *Error-Correcting Codes*, 2nd ed. MIT Press, Cambridge, Mass.
- 213. SACKS, G. E. (1958). Multiple error correction by means of parity checks. *IRE Trans. Inform. Theory* IT-4, 145-147.
- 214. SLEPIAN, D. (1956). A class of binary signaling alphabets. *Bell System Tech. J.* 35, 203-234.
- 215. SLOANE, N. J. A. (1975). *A Short Course on Error-Correcting Codes*. Springer-Verlag, Berlin.
- 216. TIETÄVÄINEN, A. On the nonexistence of perfect codes over finite fields. *SIAM J. Appl. Math.* 24, No. 1, 88-96 (1973); A short proof of the nonexistence of unknown perfect codes over $GF(q)$, $q > 2$. *Suomalais. tiedekat. toimituks, Ser. AI*, No. 580, 1-6 (1974).
- 217. TIETÄVÄINEN, A. and PERKO, A. (1971). There are no unknown perfect binary codes. *Ann. Univ. Turku, Ser. AI*, No. 148, 3-10.
- *218. VARSHAMOV, R. R. (1957). Estimate of the number of signals in error correcting codes. *Dokl. Akad. Nauk* 117, No. 5, 739-741.
- *219. ZINOVIEV, V. A. (1976). Algebraic theory of block codes correcting independent errors. *Itogi Nauki i Tekhniki, ser. Teor. Veroyatn., Matem. Statistika, Teoret. Kibernetika (Recent Results in Science and Engineering, ser. Probability Theory, Mathem. Stat., Theoretical Cybernetics)* 13, 189-234. Inst. Nauchn. i Technich. Inform. (Inst. of Scientific and Engineering Information), Moscow.
- *220. ZINOVIEV, V. A. and LEONTIEV, V. K. (1972). On perfect codes. *Probl. Peredachi Inform.* 8, No. 1, 26-35.
- 221. ZINOVIEV, V. A. and LEONTIEV, V. K. (1973). Non-existence of perfect codes over Galois fields. *Problems of Control and Information Theory (Budapest)* 2, No. 2, 16-24.

This page intentionally left blank

Name Index†

- Abramson, N. 175, 179, 193, 233, 397
 Aczél, Y. 96, 399
 Afonin, V. A. 248, 405
 Ahlswede, R. 323(fn), 406
 Air, G. M. 257, 400
 Aksakov, S. T., 201, 210
 Alekseev, P. M. 400
 Apostel, L. 209, 400
 Ash, R. B. 21, 281, 290, 306, 333, 338, 346, 397
 Ashby, W. R. xii, 397
 Atal, B. S. 197-98, 214, 405
 Attneave, C. 223, 226
 Attneave, F. 85, 191, 223, 226, 399

 Baghdady, E. J. 406
 Balasubrahmanyam, P. 197, 214, 405
 Balmont, K. D. 178(fn)
 Bar-Hillel, Y. xvii, 397
 Barell, B. C. 257, 400
 Barnard, G. A. 193, 299, 400, 405, 406
 Bartee, T. C. 306, 333, 338, 346, 406
 Basharin, G. P. 179, 197, 400
 Beauchamp, J. 224, 402
 Bektaev, K. B. 191, 195-96, 211-12, 216, 404
 Belar, H. 224, 228, 404
 Belevitch, V. 400
 Bellman, R. 120, 399
 Berg, A. I. 406
 Berlekamp, E. R. 305, 314(fn), 333, 336, 338, 346, 389, 406
 Bernstein, S. N. 42
 Berry, J. 217, 401
 Birkhoff, G. 306, 333, 338, 346, 406
 Black, J. W. 219, 401
 Bluhme, H. 197, 401
 Blyth, C. R. 179, 401
 Boltyanskii, V. G. vii(fn†)
 Boole, G. 41
 Bose, R. C. 316, 334, 341(fn), 406
 Bourbaki, N. 210

 Brawly, J. W. 224
 Brillouin, L. xii, 47, 82, 177, 397
 Brooks, F. P. 223-24, 226, 401
 Burton, N. G. 189, 191, 401

 Carnap, R. xvii, 397
 Carson, D. H. 205, 401
 Carterette, E. C. 211, 401
 Chebyshev, A. P. 33-36, 299, 306
 Chentsov, N. N. 111, 116, 400
 Cherry, Collin 177, 209, 219, 397, 404, 406
 Cherry, E. C. 220, 401
 Chiplunkar, V. N. 197-98, 214, 405
 Cohen, J. E. 224, 401
 Cover, T. M. xx, 191, 199, 201-03, 210, 401
 Crick, F. H. C. 256-57, 401
 Culbertson, J. T. 398
 Cullman, G. 305, 333, 338, 346, 397, 406

 Daroczy, Z. 96, 399
 Delbrück, M. 256, 402
 Denes, P. 219, 401
 Denis-Papin, M. 177, 397
 Deutsch, S. 238-39, 401
 Devidé, V. 120, 399
 Dewey, G. 186, 401
 Diamond, S. 4, 398
 Dmitriev V. I. 248, 405
 Dobrushin, R. L. xiv, xvii, 89, 177, 181, 281, 306, 323(fn), 324, 333, 338, 399, 401, 405, 406
 Dolezel, L. 195, 213-14, 403
 Drygas, H. 323(fn), 406

 Eidelnant, M. I. xiv
 Eldridge, R. C. 401
 Elias, P. 281, 306, 323(fn), 338, 346, 406
 Eminescu, M. 214
 Endres, W. xvii, 195, 219-20, 401

†The number appearing after each name indicates the page on which either the name or the work of the concerned individual has been referred to.

- Fadeev, D. K. 96, 399
 Fano, R. M. 147, 150, 158, 171, 175, 281, 283, 287, 306, 333, 338, 346, 397
 Faulkner, W. 213
 Feinstein, A. xii, 171, 275, 281, 299, 397
 Feller, W. 4, 398
 Fey, P. 397
 Fisher, R. A. 314(fn)
 Ford, L. R. 121, 399
 Forney, G. D. 249, 338, 406
 Forte, B. 96, 399
 Foster, Stephen 224, 228
 Foy, W. H. 245, 401
 Fradis, A. 209, 220, 401, 405
 Frick, F. C. 198, 212, 401
 Fritz, E. L. 198, 212, 401
 Frolushkin, V. C. 242-45, 401
- Gabidulin, E. M. 323(fn), 406
 Gallager, R. G. 171, 281, 290, 306, 322(fn), 324, 333, 338, 346, 397, 406
 Galois, Evariste 380(fn)
 Gamow, G. 255-57, 402
 Garmash, V. A. xiv, 194, 241, 402
 Gauss, Carl Friedrich 331(fn)
 Gel'fand, I. M. xvii, 397
 Gerstman, L. J. 196, 209-10, 214, 403
 Gilbert, E. N. 140, 299, 306, 316, 333, 338, 346, 400, 406
 Gindikin, S. G. xiv
 Giraud, R. 195, 212, 404
 Gluss, B. 120, 399
 Gnedenko, B. V. 2, 6, 398
 Golay, M. J. E. 341
 Golomb, S. W. 256, 402
 Goncharov, I. A. 201, 210, 213
 Goryachaya, M. M. ix
 Gorenstein, D. C. 342, 406
 Grier, G. W. 198, 212, 401
 Griffith, J. S. 256, 401
 Grignetti, M. 209, 402
 Gunzenhäuser, R. 213, 402
- Halle, M. 220, 401
 Halmos, P. R. 388(fn)
 Hamming, R. W. 314-17, 326, 406
 Hansson, H. 195, 403
 Hardy, G. H. 305(fn)
 Hartley, R. V. L. 53-56, 59, 125, 147
 Haton, J. P. 219, 402
 Haydn, J. 224
- Hick, W. E. 83, 399
 Hiller, L. 224, 402
 Hintikka, J. xvii, 397
 Hockquenghem, A. 305, 316, 333-34, 338, 346, 406
 Hodges, S. L. 398
 Hopkins, A. L. 223-24, 226, 401
 Huffman, D. A. 153, 155-57, 171-72, 177, 259, 400
 Hutchison, C. A. 257, 400
 Hyman, R. 57, 73, 399
- Ibragimov, T. I. 195, 402
 Inagaki, Y. 407
 Ivadari, E. 407
 Ivanov, V. V. xvii
- Jacobs, I. M. 247-49, 306, 333, 338, 346, 398
 Jacobson, H. 250, 402
 Jacobson, R. 220, 401
 James, William 211
 Jamison, D. 402
 Jamison, K. 402
 Jensen, J. L. W. V. 352
 Johnson, S. M. 121, 399
 Jones, M. H. 211, 401
 Joyce, J. 209, 213
- Kac, M. 2, 4, 398
 Kaplan, A. J. 237, 405
 Karush, J. L. 175, 400
 Kasami, T. 407
 Kaufmann, A. 306, 333, 338, 346, 407
 Kayser, G. A. 191, 220, 239-41, 402
 Kazaryan, R. A. 194-95, 402
 Kellogg, D. J. 120, 399
 Kellogg, P. J. 120, 399
 Kelly, D. H. 250, 402
 Kemeny, J. G. 398
 Kharkevich, A. A. xiv, 247-48, 402
 Khinchin, A. Ya. 2, 6, 398
 Khorana, H. G. 257
 King, R. C. 191, 199, 201-03, 210, 401
 Kirillov, N. E. 241, 402
 Kislitsin, S. S. 120-21, 400
 Kolesnik, V. D. 333, 338, 406
 Kolmogorov, A. N. ix, xiv, xvii, 2, 4, 43, 89, 191, 195, 198-99, 201-02, 210, 212-15, 397, 398
 Kondratov, A. M. 213, 402

- Kordemskii, B. A. 101, 104, 108, 111, 400
 Kotel'nikov, V. A. 247
 Kraft, L. G. 175
 Kreindler, A. 220, 405
 Kreuzer, H. 213, 402
 Krinchik, E. P. 85, 399
 Kuiper, N. H. 388(fn)
 Küpfmüller, K. 195, 218-19, 241, 250, 402
 Kuprin, A. I. 214

 Lamotte, M. 219, 402
 Laplace, P. S. 43
 Lebedev, D. S. xiv, 194, 233, 235-36, 402, 403
 Leder, Philip 257
 Lehmann, E. L. 398
 Lenskii, D. N. 195, 402
 Leonard, J. A. 85, 399
 Leonard, Zunin 202
 Leontiev, A. N. 85, 407
 Leontiev, V. K. 342, 399
 Levenstein, V. I. xiv
 Levin, L. A. xvii, 398
 Levinson, N. 305(fn††), 306, 333, 338, 346, 407
 Licklider, J. C. R. 189, 191, 401
 Limb, J. O. 233-36, 403
 Lin, S. 305, 333, 338, 346, 407
 Lint, van J. H. 305, 333, 338, 346, 407
 Lomov, B. F. 85, 251, 399
 Luce, R. D. 85, 399
 Lüdtke, H. 213

 MacWilliams, F. J. 305, 333, 338, 346, 407
 Maixner, V. 190, 403
 Malone, D. 202, 210
 Mandelbrot, B. xx, 209, 217, 400, 403
 Manfrino, R. 193-95, 203, 209-10, 403
 Marcus, S. 213-14, 403
 Matthaei, J. H. 257
 McMillan, B. 171, 175, 400
 Meshalkin, L. D. 399
 Michel, S. W. 241, 245, 403
 Mihailescu, L. 209, 220, 401, 405
 Miller, G. A. 179, 209, 216, 403
 Minkowski, H. 388(fn)
 Mirkil, H. 398
 Mironchikov, E. T. 333, 338, 407
 Moles, A. 398
 Moloshnaya, T. N. xiv
 Moore, E. F. 140, 400
 Morf, A. 209, 400
 Mosteller, F. 2, 4, 167, 399

 Nair, K. K. 197-98, 214, 405
 Neidhardt, P. 235, 238, 248, 403
 Nemetz, T. xx, 179, 192, 203, 403
 Neumann, P. G. 223-24, 226, 401
 Nevel'skii, P. B. 191, 210, 403
 Newman, E. B. 196, 209-11, 214, 250, 403, 404
 Neyman, J. 4, 399
 Ng, C. T. 399
 Nicolau, E. 195, 213-14, 403
 Nikolaev, V. I. 85, 399
 Nirenberg, M. W. 257, 401
 Novikov, P. S. xiv

 Ochoa, S. 257
 Olson, H. 224, 228, 404
 Orgel, L. E. 256, 401
 Ovseevich, I. A. xiv, xvii
 Ozhegov, S. I. 215

 Paisley, W. J. 214, 404
 Parkhomenko, P. P. 122, 400
 Parks, J. R. 236, 404
 Patterson, G. W. 140, 400
 Perko, A. 342, 407
 Peshkovskii, A. M. 220, 404
 Peterson, W. W. 305, 333, 336-38, 342, 346, 406, 407
 Petrova, N. V. xvii, 194-95, 211, 404
 Pfaffelhuber, E. 179, 404
 Picard, K. 122, 400
 Pierce, J. R. 178, 209, 224, 228, 398
 Piil, E. I. 233, 235-36, 403
 Pinkerton, R. C. 222-23, 225-26, 404
 Piotrovskaya, A. A. 195-96, 211-12, 216, 404
 Piotrovskii, R. G. 177, 195-96, 211-12, 216, 404
 Platzman, R. L. 398, 402
 Poleyatev, I. A. 47, 398
 Pratt, F. 179, 182-83, 186-87, 194, 205, 404
 Prokhorov, A. V. xvii
 Pushkin, A. S. 213, 215

 Quastler, H. 59, 85, 178, 250, 398, 402, 404

 Rajagopalan, K. R. 191, 197, 214, 224, 405
 Ramachandran, V. 197-98, 214, 405
 Ramakrishna, B. S. 197-98, 214, 404, 405
 Ray-Chaudhuri, D. K. 316, 334, 406
 Razzhivin, K. A. 211, 404
 Reza, F. M. 179, 404

- Rich, A. 257, 402
 Roćeric, A. 195, 213-14, 403
 Roland, M. 224, 404
 Rosenbaum, M. D. 191, 210, 403
 Rourke, R. E. K. 2, 167, 399
 Rychkova, N. 199, 404
 Rytov, S. M. xiv

 Sacks, G. E. 326, 407
 Sala, C. 195, 213-14, 403
 Sardinias, A. A. 140, 400
 Savchuk, A. P. 190, 404
 Schober, H. 250, 404
 Schőnberg, A. 224
 Schreiber, W. F. 231-33, 235-36, 404
 Schreider, Yu. A. xvii, 398
 Schubert, F. 224
 Schumann, R. 224
 Shaker, M. M. 194, 403
 Shannon, Betti 210
 Shannon, C. E. vii, viii, xii, xviii, 53-56, 96, 147ff., 150, 158, 171, 180-81, 186-92, 195, 197-98, 201-02, 204, 209-10, 212, 215, 236-37, 247-48, 250, 259, 270, 274-75, 281-83, 291, 301, 305, 314(fn), 333, 398, 404, 406
 Shaw, Bernard 217
 Shestopal, G. A. xiv
 Shklarsky, D. O. 111, 116, 400
 Simon, J. 192, 203, 403
 Siromoney, G. C. 191, 197-98, 214, 224, 404, 405
 Slepian, D. 299, 306, 323, 333, 338, 346, 406, 407
 Sloane, N. J. A. 305, 333, 338, 346, 407
 Smirnov, O. L. 212, 405
 Snell, J. L. 398
 Spencer, R. V. 237, 405
 Stambler, S. Z. xvii
 Steinhaus, H. 120, 400
 Stevens, S. S. 216, 403
 Subramanian, R. 197-98, 214, 404-405
 Sumby, W. H. 198, 212, 401
 Suppes, P. xvii, 397
 Sziklai, G. C. 250, 405

 Taha, E. H. 194, 403
 Tarnóczy, T. 213-14, 405
 Temnikov, F. E. 248, 405
 Thomas, G. B. 2, 167, 399

 Thompson, G. L. 398
 Thorndike, E. L. 59, 186, 405
 Tietäväinen, A. 342, 407
 Tokura, N. 407
 Tolstoy, L. N. 194
 Tsannes, N. S. 237, 405
 Tsukkerman, I. I. 233, 235-36, 403
 Tsybakov, B. S. xvii

 Urbach, V. Yu. 187, 405
 Uspenski, V. A. xiv, 219, 405

 Varshamov, R. R. 316, 407
 Vasiliev, R. R. 242-43, 405
 Voinescu, I. 209, 220, 401, 405

 Wanas, M. A. 194, 403
 Waugh, N. C. 196, 210-11, 214, 404
 Webern, A. 224
 Welch, L. R. 256, 402
 Weldon, E. Y. 305, 333, 336-38, 342, 346, 407
 Welford, A. T. 85, 399
 Weltner, K. 177, 191-92, 211, 405
 Wiener, N. 398
 Wolfowitz, J. 281, 290, 398
 Woodward, P. O. xii, 247-48, 398
 Wozencraft, S. M. 247-49, 306, 333, 338, 346, 398
 Wright, Earnest Vincent 178
 Wright, W. V. 223-24, 226, 401

 Yaglom A. M. xvii, 15, 42, 177, 397, 399, 405
 Yaglom, I. M. 15, 42, 111, 116, 177, 399, 400, 405
 Yćas, M. 256-58, 402, 405
 Yekimov, A. V. 212, 405
 Youngblood, J. E. 224, 405
 Yockey, H. P. 398, 402

 Zaidman, R. A. 215
 Zaremba, S. K. 290-91, 406
 Zaripov, R. Kh. 224, 228, 405
 Zayed, A. I. 194, 403
 Zierler, N. 342, 406
 Zinder, L. R. xiv, 405
 Zinoviev, V. A. 306, 333, 338, 342, 346, 407
 Zipf, G. K. 209, 405
 Zitek, F. 403
 Zvonkin, A. K. xvii, 398

Subject Index

- Absolute value (or norm) of
 - elements 41-42, 388
 - numbers 41-42
- Adenine 253, 255
- Aitham 198
- Algebra, Boolean 41-43
 - link with probability theory 42
 - normed 41-43
- Algebra of
 - events 36ff., 37, 40
 - numbers 37
 - sets 37
- Algebraic
 - coding theory 328-29
 - concepts xvi, 364-91
 - operations 364
 - system 37, 41, 365
- Algorithm, Euclidean 376
- Algorithmic approach to the
 - concept of amount of information xvii
- Alphabet 139, 252 (*see also* under various languages)
 - contraction of 153-57, 172
 - one-fold 154
 - two-fold 154
 - inclusive of space 203-06
 - spaceless 203-06
- Amino-acid 253, 255
 - proline 257
- Amusing problems (*see* Recreative problems)
- Aphasic persons' speech, entropy of 210, 221
- Arabic alphabet 194
- Arithmetic
 - q - 365-66, 368
 - Qx - 378
 - with two symbols (2-arithmetic) 355, 370-71
- Arithmetic mean 31-35, 290, 297, 349-50
 - theorem 355
- Bacteria 253
- Baudot code 138, 140
- Binary channel (*see* under Communication channel)
- Binary code (*see* under Code)
- Binary field 317
- Binary number system 143, 145, 307
- Binary system of logarithms xiii, xv(fn)
- Binary unit (bit) 45, 144
- Block codes (*see* under Code)
- Block, N -letter 162, 289
- Boolean algebra 41-43
 - link with probability theory 42
 - normed 41-43
- Brightness levels 230-33
- Centre of gravity 353
- Centroid 353
- Channel capacity (*see* under Communication channel)
- Chebyshev's inequality 26f., 33-36
- Chromatic scale 222-23
- Chromosome structure 253
 - double helix 253
- Code(s) 138, 141, 338
 - advantageous (efficient) 137f., 190
- Baudot 138, 140
- binary 138, 141-42, 148-49
 - Golay perfect 341
 - most efficient, construction of 142
 - non-uniform 140
 - uniform, most efficient 144
- block 145, 313(fn)
- Bose-Chaudhuri-Hocquenghem xx, 316, 334-37, 345(fn), 346
 - check matrix of 335
 - (N, M) - 344
 - non-primitive 324(fn), 341(fn), 342
 - primitive 324(fn)
- check matrix of 321, 323-24, 327-28, 345
- composition 256
- correction by check signal 310
- cyclic 331, 334, 344
 - polynomial of 333
- decimal 174
- degenerate 257
- densely-packed 339-40

Code(s) (*contd.*)

distance 338
 double-error correcting 314, 326-27, 334-35, 344
 error-detecting and error-correcting 304-46
 definition 308
 Fano 150
 efficiency (economy) of 137-47
 generated by a polynomial 329-31
 genetic 255-58
 Gamow's postulation 255
 structure of 257
 without comma 257
 Golay 341
 group 319-21, 339
 Hamming 314-15, 325-26, 333, 341-42, 345
 extended 326
 Huffman xvi, 147f., 153-57, 171-72, 259, 313(fn)
 optimality of 156, 172, 177
 ternary, construction of 172
 instantaneously decipherable 141, 157, 175-77
 linear 319-21, 339
 m -ary 147, 172-73, 175
 Morse 138, 140
 (N, M) - 313-14, 344
 non-uniform 140-41, 143
 optimal 156-58, 336
 overlapping 255-56
 parity-check 310, 312-13, 315-16, 318, 321, 323, 333
 decoding of 322
 matrix of 319
 (N, M) - 317-18
 non-random 324
 random 323
 systematic 319
 perfect 339-42
 polynomial 329, 332
 ideal in 332
 quasi-perfect 342
 redundancy 308-09
 $(7, 4)$ -single error correcting 310-11, 314
 Shannon-Fano xvi, 147, 150-53, 171, 259, 313(fn)
 single-error-correcting 309-14, 327, 335, 344, 345(fn)
 systematic 319
 ternary 138
 triple-error-correcting 335, 346
 triple repetition 308-09, 313(fn)

Code(s) (*contd.*)

triplet 257
 uniform 138, 140-41, 143, 148, 166
 uniquely decipherable 140, 150, 175-77
 with/without separating symbol, comma 140, 256
 Code-word 140, 144, 155, 159-60, 166, 175
 choice of, simple method for 306-07
 length 144
 average 151, 156-57
 Coding 122, 138, 249, 251-53, 255
 and statistical laws 148
 advantageous (efficient) 139
 algebraic theory of 317(fn), 318
 block 146, 152-53, 157, 289
 fundamental theorem of xvi, 147f., 157-63, 172-73, 246, 274
 fundamental theorem of noisy (due to Shannon) xvi, 274-75, 289-306
 converse of xvi, 273, 283-90
 strongly converse of 290
 Huffman method of 155-57
 method 293, 307
 random 292-97, 301(fn), 306, 323
 Shannon-Fano method of 152-53, 155-56, 158
 theory xvi, 305-07
 notion of entropy in 161
 Codon 255-56, 258
 'Comma', as separating symbol 140
 Communication, specific content of 55
 Communication channel xii, xiii
 and statistical regularities 55
 associated with hereditary phenomenon 252
 binary
 asymmetric 273
 symmetric 263, 265, 275, 277, 280, 299, 301(fn), 302, 307-308, 311, 323, 342
 erasure 267, 269-70, 323
 capacity xiii, 246-51
 in absence of noise 173, 262
 in presence of noise 262-63, 272, 297
 zero-error 283
 human organism as 249-51
 m -ary symmetric 266
 new forms of 249
 noisy 252, 260-62
 code selection for every 274
 mathematical model of 260-61
 non-binary 317(fn), 346
 transmission of speech over 219

- Complement of a set 39
- Constant
 - mean value of a 28
 - number p 1
- Convex k -gon 354
- Coset 368
- Counterfeit coin problems viii, 108-21, 136, 147
- Current pulse 137, 139, 246, 251
- Cytoplasm 252-53
- Cytosine 253, 255
- Czech language
 - entropy of 214
- Decimal number system 143
- Decimal system of logarithms xiii, xv
- Decimal unit (dit) xv, 46
- Decoding 140-41, 249, 251-53, 255
- Decoding error probability 293, 296
 - mean 339-40, 342
 - Hamming lower bound on 340
- Decoding, instantaneous 141
- Decoding method 293, 299, 306-07
- Decoding rule 322
- Decoding, sequential 251, 323-24
- Decoding, unique 140
- Deoxyribonucleic acid (DNA) 253-54
 - molecules, four-letter alphabet 253
- Die
 - imperfect 42
 - throw 1-2
- Disjoint equal spheres, closest packing of 389
- Dispersion (or spread) 27-34, 36 (*see also* Variance; Variance of random variables)
- Distance 387
 - code 338
 - Euclidean 389
 - geometric 338
 - Hamming 338, 340, 342, 389(fn)
 - utility in coding theory 389
 - Lee 389
 - Minkowski 388
- Distributive law 38
- Dit xv, 46
- Divisor of numbers, greatest common 40, 365, 375
- Dravidian languages, entropy of 197, 214
- Ear, resolving power of 247
- Element(s)
 - difference of 365
 - identity 364
 - inverse of an 365
 - symmetric 364-65
 - unit 365, 371
- English alphabets 139, 192, 203
- English language
 - average information in stressed words of 217
 - average word-length in 181, 188
 - coding text-letters in, Shannon-Fano method 180
 - digram frequency in 182-83, 186
 - entropy of letters/words in 179-81, 187, 194
 - estimation by Cover and King 202-03
 - first-order approximation to 181, 186
 - letter/word frequency in 178-79, 182, 186
 - letter-guessing experiments for 191, 196
 - redundancy for 185, 188, 191, 195-96
 - second-order approximation to 183, 186-87
 - statistical characteristics of letters/words 186-88
 - third-order approximation to 183
 - trigram frequency in 183, 186
 - zero-order approximation to 178
- English speech
 - one phoneme redundancy for 220
- Entropy(ies) viii, xvii, 44ff., 47, 53-54, 74, 88, 93f.
 - addition law of 60, 63, 98
 - as measure of uncertainty 44ff., 47, 56, 94
 - combinatorial 215
 - concept of xii, 55-56
 - conditional 62-63, 181-82, 184, 188-90, 220
 - ϵ - 82, 230
 - Hartley's viewpoint of 53-56, 59, 125
 - limit 185, 188-89, 207
 - of compound experiment 61
 - of experiment 47
 - of language 177f. (*see also* under Various languages)
 - method of determining 198
 - of N -letter block 162
 - of one letter 149
 - of speech 215f.
 - of television images 231
 - residual 124-25
 - Shannon's approach to 53-57, 215
 - specific 162, 171, 184, 198, 219
 - thermodynamic concept of 47(fn.†)
 - true, upper bound on 199
 - unconditional 71

- Enzymes 253
- Equiprobability 3-4, 44-45
- Error(s)
 - single 309, 311
 - systematic 26, 32
- Error probability
 - exponential bound of 281(fn)
 - mean value of 284-86
- Event(s)
 - addition of 7f., 36
 - certain (sure) 8
 - compatible 9
 - contrary 8
 - incompatible 7f.
 - pairwise 9, 23, 45, 59
 - impossible 8, 35
 - independent 7f., 10-11, 20, 25
 - multiplication of 7f., 36
 - associative law of 38, 91
 - obeying statistical law 82
 - product of 10, 36
 - random 1f., 4, 40, 44
 - set of elementary 43
 - sum of 8-9, 14, 36
- Euclidean
 - distance 389
 - division of algorithm 376(fn)
 - ring 376(fn)
 - space 387
- European languages
 - entropy values for 198
 - letter frequencies in 197
 - redundancy estimates for 185, 196
 - spoken, phoneme statistics and entropies 219
- Experiment(s) 1, 121-22 (*see also* under Letter-guessing; Urn)
 - auxiliary 1, 122
 - compound 42, 45, 55, 88-89, 123, 125, 232
 - dependent 61
 - equally probable outcomes of an 42-43, 45
 - independent 45, 59
 - simple 88
 - weather 52
- Exponential mean 349-50
- Eye, resolving power of 231, 247
- Fano codes (*see* under Codes)
- Fano inequality 287-88, 303-04
- Field 369-71, 379
 - algebraic, two-element 328
 - basic (or field of scalars) 382
 - binary 317
 - finite 317(fn)
 - Galois' 380(fn)
 - order of 371
 - primitive element of 372
 - properties of 371
- Formula for the number $\binom{N}{K}$ 14-15
- French language
 - digram probabilities in 194
 - entropy of 195
 - first-order approximation to 193
 - letter frequency in 192-93
 - letter-guessing experiments for 196
 - phoneme statistics and digrams of 219
 - redundancy estimates for 195
 - trigram probabilities in 194
 - word-length in 192
- Frequencies, limiting, of guessing a letter 189
- Frequency
 - dictionary 186, 207, 217
 - of occurrence of result 1
- Function(s)
 - convex 287, 304, 347ff.
 - graph of 351
 - properties of 347ff.
 - test for 346(fn)
 - exponential 281(fn), 347
 - logarithmic 347
- Gambling scheme, due Cover and King 201-03
- Game of 'garbled telephone' 89
- Gene, 'initiation' and 'termination' marks 256(fn), 257
- Genetic information transmission (*see* under Information)
- Geometric distance 338
- Geometric mean 349
 - theorem 304, 355
- Geometry
 - discrete 389
 - Euclidean 389
- German language
 - 'first-order approximation' to 193
 - letter frequency in 192-93
 - letter-guessing experiments for 195-96
 - redundancy for 195
 - 'semantic' information in, study by Küpfmüller 218

- German speech
 - redundancy of one phoneme 220
- Golay perfect binary code 341
- Greatest common divisor 40, 365, 375
- Group 320, 364-65
 - commutative 364
 - cyclic 369
 - multiplicative 365
 - non-commutative 364(fn)
 - null element of 365
 - order of 369
 - subgroup of a 320, 368
- Guanine 253, 255
- Guessing method viii, 148-49, 181, 188-92, 195-98, 204, 206, 211, 225
 - due Shannon 188-92, 195-98, 201-02, 212, 236-37
 - sharpening by Kolmogorov 198-99, 212, 214
- Hamming
 - distance 338, 340, 342, 389(fn)
 - inequality 315, 328, 335-37, 339, 341
 - lower bound 315-16, 336-37, 339
 - metric 388-89
 - utility in coding theory 389
 - sphere 338, 340-42, 389
 - upper bound 339
- Hebrew language
 - redundancy for 197
 - statistical structure of 197
- Hieroglyphic writing 207-08, 231
- Hungarian
 - alphabets 203
 - language, entropy of 203
 - prose, information-theoretic characteristics of 214
 - redundancy estimates for 203
- Ideal 332, 376
 - principal 377, 380
- Images
 - colour television 237-38
 - detail-starved 235
 - heterochromatic 235
 - monochromatic 232, 235
 - television 288f.
- Indian languages 197-98, 214
 - redundancy estimates for 197-98
- Inequality
 - Chebyshev 26f., 33-34, 299, 306
 - Fano 287-88, 303-04
 - Hamming 315, 328, 335-37, 339, 341
 - Jensen 352, 356
 - Kraft 175
 - McMillan 175
 - Varshamov-Gilbert 316, 327-28, 335-37, 345
- Information
 - amount of xiii, 74, 215
 - algorithmic approach to xvii
 - combinatorial approach to xvii, 215
 - average amount of
 - in an experiment 74-75, 132
 - in a text-letter 203
 - in a word 207
 - concept of 7, 73f., 80, 90, 137
 - conditional 91-92
 - entropy and 44ff., 74
 - genetic, and its transmission 251-58
 - limiting 214
 - mean 74-75
 - conditional 91
 - of various messages encountered in practice 177ff.
 - as continuously varying messages (television images) 228f.
 - as musical notes 222f.
 - as phototelegrams 238f
 - as spoken language 215f.
 - as written language 177f.
 - reciprocal, of two events 86
 - semantic xvii, 216, 218-19, 221(fn), 228-29
 - specific 205, 207
 - symmetry property of 91, 93
 - total 205-06, 221
 - trip'e, equation 92-93, 299
 - unsemantic 218-19, 229, 231, 250
- Information theory vii-viii, xi-xiii, xv-xvii, xix
 - applications to information transmission through communication channels viii, 137ff.
- Information transmission
 - error-free 282, 283(fn)
 - over noisy channels 258-304
 - rate 173, 216(fn), 304-05
 - largest 339
 - sequential 89
- Input signals 251, 254-55
- Insistence stresses 217
- Inverse of an element 365

- Italian language
 - entropy of 209-10
 - letter frequency in 193-94
- Jensen's inequality 352
 - general 356
- k -arithmetic 368
- Kazakhian language, letter-guessing experiments for 196
- Kraft's inequality 175
- Lagrange's theorem 369
- Language (*see* under Entropy; Redundancy; also Various languages)
 - control tower 212
 - intonation 218
 - specialized 210, 212
 - spoken xiii, 215f.
 - statistical structure of 188
 - written xiii, 177f.
- Law of
 - contradiction 41
 - excluded middle 41
 - large numbers xvi, 26f., 36, 170
- Least common multiple 40
- Lee distance (metric) 389
- Letter(s), average frequency of 178
- Letter-guessing experiment (*see* also Guessing method)
 - by Piotrovskii *et al.* 211
 - by Weltner 211
 - technique
 - Kolmogorov's postulate on 201-02
 - refinement by Cover and King 201
- Linear
 - code-word collection 320
 - subspace 320-21
- Logarithm
 - decimal 46
 - factor of transition $\log_b a$ 45
- Logic, mathematical 41-42
- Logical
 - problems 100-08 (*see* also Recreative problems)
 - on geometric probability 42
 - propositions 41-42
- Luminosity scales (*see* Brightness levels)
- m -ary number system 143, 147
- Mathematical logic 41-42
- Matrix (Matrices) 318, 366
 - additive group of 366
 - check 321
 - elementary transformations on 390
 - equivalent 390
 - multiplication 389
- McMillan inequality 175
- Mean 6, 26
- Mean value of
 - square of deviation 27
 - weighings 133, 136
- Melody (Melodies) 218
 - computer-created 226-28
 - new, by urn experiments 225
 - redundancy for 223
 - Stephen Foster's 224, 228
- Mesh of an element block 243-44
 - value calculation by Frolushkin 243
- Message(s) xiii, 136, 251 (*see* also under Information)
 - continuously varying, television images 228-38
 - entropy of 236-37
 - redundancy for 236-37
 - discrete, optimal coding/decoding of 252
 - input 253-55
 - musical (*see* under Musical)
 - output 253-54
 - phototelegraphic xiii
 - entropy of 238-39, 242
 - redundancy for 242
 - screen elements of 238-39, 241
 - specific entropy of 243
 - statistical laws of xiii, 149
- Message quantization 230
 - probable sequences of 171
- Message transmission
 - error-free 273, 283
 - mean error frequency in 284(fn)
 - rate 246-50, 273
- Metric
 - Hamming 388-89
 - utility in coding theory 389
 - Lee 389
 - Minkowski 388
- Metric space 387
- Minkowski metric 388
- Mitosis 253
- Monte-Carlo methods 226

- Morpheme 208
- Morse code 138, 140
- Music 222f.
 - basic elements 222-23
 - chromatic scale 222-23
- Musical compositions
 - cowboy songs, American 223, 226
 - entropy of, maximum possible 222
 - German romantic 224
 - Haydn's 224
 - hymns 223-24, 228
 - information-theoretic characteristics 224
 - note-guessing method due Pinkerton 225
 - nursery-rhymes, American 222
 - redundancy of 222, 224-26
 - rock and roll, American 224
 - Schönberg's 224
- Musical messages 222
- Musical sentences 225
- Mutation 253

- Nerve cells (neurons) 252
- Nerve fibres
 - aural 250
 - optic 250, 252
- Non-parity 314
- Number(s) 366
 - absolute value (norm) of 27, 41-42
 - algebra of 37
 - comparison of, basic rules 39
 - digits 143-44
 - greatest common divisor of 40, 365, 375
 - idempotent 37
 - law of large xvi, 26f., 36, 170
 - least common multiple of 40
- Number system
 - binary 143
 - centenary 145
 - decimal 143
 - m*-ary 143
 - ternary 147

- Order of
 - field 371
 - group 369
 - ring 377-78
- Oriental languages 197
- Outcomes 4
 - equally probable 3-4, 130-31
 - complete system of 42-43
 - impossible 55
 - low-probability 59
 - mutually exclusive 9
 - not equally probable 131
 - proof of feasibility/infeasibility of 128
- Output signal 254-55

- Parity 314
- Parity check 310, 312-13, 315, 318
 - matrix of a code 318
- Pause 137
- Phenylalanin 257
- Phoneme(s) 208, 219, 229
 - alphabet 219
 - average length of 221-22
 - spectrogram 220
- Phototelegrams 238f.
 - message transmission by 238-51
 - screen elements employed in 238
- Phrase 208
- Pitch of tone 229
- Polish language, letter guessing experiments
 - for 196
- Polynomial(s) 367
 - as a group 367
 - check 341
 - code generator 329, 342-43
 - composite 379
 - cyclotomic 331(f_n)
 - irreducible 379
 - minimal 343
 - reducible 381
 - roots of 342
- Portuguese language
 - entropy of 193, 210
 - letter frequency in 193
- Probabilistic choices reduced to binary choices 225
- Probability(ies) 1ff.
 - addition law of 9, 87
 - basic 42
 - classical concept of viii, 42-43
 - conditional 20f., 61
 - definition 1f., 42
 - mean error 283-84, 286-87, 288(f_n)
 - multiplication law of 11, 22, 25
 - properties of 7f.
 - table 4
 - total, equation of 23
- Probability theory viii, 2, 42-43
 - an approach first indicated by Bernstein 42
 - axiomatic construction due Kolmogorov 43

- Probability theory (*contd.*)**
 definition, classical, introduced by Laplace 43
 link with Boolean algebra 42
 main problem of 42
Problems
 counterfeit coin 108-21
 logical 100-08
 on geometric probability 42
Protein(s) 253-54
 synthesis 254
 twenty-letter language 255

 q -arithmetic 365-66, 368
Quantization of messages 230
Question(s) 122
 cost element of 122
 mean value of 130, 134
Questionnaire 122
 $Q(x)$ -arithmetic 378

Random event 1f.
Random variable(s) 1f., 170, 293
 arithmetic mean of 31-32, 34-36
 independent 16-17, 29
 mean value of 6, 26, 54
 any 296-97
 product of 19
 sum of 16, 18, 29, 31
 mutually independent 19, 31, 36
 pairwise independent 36
 product of 15, 18-19
 standard deviation of 27
 sum of 15-16, 18, 29-31
 variance of 26f.
 square root of 27
Reaction
 complex 57
 error-free 83
 of choice 57
 psychic 56
 simple 57
 time, mean 57-58, 72-73, 83-85
Reciprocal information of two events 86
Recreative problems viii, xii, xix, 101, 106, 108, 137
Redundancy, zero 188, 213
Redundancy estimates for
 a language 185, 188, 190-91 (*see also under*
 Various languages)
 business texts 211
 literary texts 210-15
 melody 223
 speech 216
 relationship with that of written language 221
 television images 231-37
 typed text 240-41, 244
Ribonucleic acid (RNA) 254
 messenger (mRNA) 254-55
 molecule, information transmission through 254-55
Ribosomes 253-54
Ring 369, 374
 commutative 369, 374(fn)
 Euclidean 376(fn)
 field as a 375
 order of 377-78
Rumanian language
 entropy estimates for 209-10, 214
 letter-guessing experiments for 196
Rumanian speech, low-order entropy estimates for 220-21
Russian alphabets 139, 194
Russian language
 letter-guessing experiments for 196
 redundancy estimates for 196-201
Russian literary texts, guessing method for 199-201
Russian speech, entropy of phonemes in 220
Russian telegraphic text 212
Russian tetrametric iambic verse 213

Samoyan language, entropy and redundancy estimates for 196
'Saturation' of a block of elements 243-44
 calculation of values of, by Frolushkin 243
Semantic information xvii, 216, 218-19, 221(fn), 228-29
Sense organs, information receiving capacity of 250-51
Set(s)
 comparison of 39
 complement of a 39-40
 empty 38
 intersection of 38, 40
 ordering of 40
 product of 38
 sum of 37
 super- 43
 union of 37, 40, 90

- Shannon's approach (*see* under Entropy)
- Shannon's coding theorem (*see* under Coding)
- Signal(s) 137, 251
 - binary 149, 157
 - check 310, 312-13, 315, 336, 345
 - distortion of 247-48, 258-59
 - elementary 138-39, 146-48, 151, 155, 251, 254, 285(fn)
 - average number of 142, 147-48, 155-56
 - error-free transmission of 265
 - information 310, 312-13, 336, 345
 - maximal level of 248
 - probabilistic average value of 148
- Sorting of objects 122
- Space
 - Euclidean 387
 - metric 387
 - subspace, linear (or vector) 320-21, 384-85
 - vector 319, 381
 - dimension of 382
- Spanish language
 - 'first-order approximation' to 193
 - letter frequency in 193
- Speech melodies (*see* under Melody)
- Sphere
 - disjoint, equal, closest packing of 389
 - Hamming 338, 340-42, 389
- Statistical regularities, role of, in communication lines 55
- Subgroup
 - cyclic 369
 - index of 369
- Subspace linear (or vector) 320-21, 384-85
- Swedish language, redundancy estimates for 195
- Syllable 208
- Systematic errors 26, 32
- Tamil alphabets 197-98
- Television images xiii (*see* under Entropy, also Redundancy)
- Ternary number system 147
- Tetrahedron, throw of 25, 90, 93
- Text words, statistical relationship between 207
- Thesaurus xvii
- Thorndike dictionary 59
- Thymine 253-55
- Total probability equation 360
- Transformations, elementary (*see* under Matrix)
- Transmission band width 247
- Trials 1
 - mutually independent 31
- Triangle inequality 387
- Uncertainty
 - degree of 44-45, 56-57
 - entropy as a measure of the amount of 44f.
 - mean value of 54
 - measurement (*see* also under Entropy)
 - Hartley's viewpoint 53, 125
 - in binary unit (bit) 45-46
 - in decimal unit (dit) xv, 46
 - Shannon's approach to 53
- Unit(s)
 - binary (bit) 45-46
 - cube 389
 - decimal (dit) xv, 46
- Upper bound
 - Hamming 339
 - Varshamov-Gilbert 316, 336-37
- Uracil 254-55, 257
- Urn problem(s) xii, 2-3, 42, 51
- Variables, random (*see* Random variables)
- Variance 27-34, 36
- Varshamov-Gilbert
 - inequality 316, 327-28, 335-37, 345
 - upper bound 316, 336-37
- Vector(s)
 - coordinates of a 342
 - linearly dependent 386
 - linearly independent 386
 - null 366
- Vector-column 366
- Vector-row 366
- Vector space 319, 381
 - dimension of 382
- Weather
 - experiments 52
 - forecast 53, 77
- Words
 - as blocks 208
 - space between 203
- Zipf principle 209

